Colecția **Economikon** este coordonată de prof.univ.dr. Ion Pohoață.
Colecție fondată de prof.univ.dr. Dumitru Zaiț și prof.univ.dr. Ion Pohoață.

Dănuț-Vasile Jemna este profesor universitar la Facultatea de Economie și Administrarea Afacerilor, Universitatea „Alexandru Ioan Cuza" din Iași. În cadrul specializării de Statistică și Previziune Economică, predă cursurile de Teoria și practica sondajelor, Econometrie și Demografie, la nivel de licență, iar la programele de master predă Econometrie avansată, Metodologia cercetării și Statistică bayesiană. A realizat stagii de cercetare doctorală și postdoctorală la Universitatea Poitiers din Franța și a realizat numeroase mobilități de predare în cadrul programelor Erasmus în țări precum Italia, Franța, Polonia. Este coordonator de doctorat în domeniul *Cibernetică și Statistică* și visiting profesor al Universității din Poitiers, Franța.

Referenți științifici:
Prof. univ. dr. Carmen Pintilescu, Universitatea „Alexandru Ioan Cuza" din Iași
Prof. univ. dr. Dorina Lazăr, Universitatea „Babeș-Bolyai", Cluj-Napoca

Redactor: Cerasela Cirimpei
Coperta: Manuela Oboroceanu

ISBN 978-606-714-482-6

© Editura Universității „Alexandru Ioan Cuza", 2018
700109 – Iași, str. Pinului, nr. 1A, tel./fax: (0232) 314947
http:// www.editura.uaic.ro e-mail: editura@uaic.ro

Sondajul statistic

DĂNUȚ-VASILE JEMNA

Ediția a II-a revizuită și adăugită

Editura Universității „Alexandru Ioan Cuza" Iași

2018

Descrierea CIP a Bibliotecii Naţionale a României
JEMNA, DĂNUŢ
Sondajul statistic / Dănuţ-Vasile Jemna. - Ed. reviz. şi adăug. -
Iaşi: Editura Universităţii „Al. I. Cuza", 2018
Conţine bibliografie
ISBN 978-606-714-482-6

311

Cuprins

INTRODUCERE 9

1. FUNDAMENTELE CERCETĂRII PRIN SONDAJ 11
1.1. DE LA CERCETĂRILE EXHAUSTIVE LA CELE SELECTIVE 12
1.1.1. Terminologie 12
1.1.2. Scurt istoric şi condiţii pentru dezvoltarea cercetării prin sondaj 13
1.1.3. Opţiunea între sondaj şi cercetarea exhaustivă 16
1.1.4. Limite ale utilizării sondajelor 17
1.2. CONCEPTUL DE REPREZENTATIVITATE A UNUI SONDAJ 19
1.2.1. Elaborări empirice 19
1.2.2. Elaborări ştiinţifice 21
1.3. PROBLEME TEORETICE ŞI METODOLOGICE 24
1.3.1. Definirea temei de cercetare 26
1.3.2. Elaborarea chestionarului statistic 39
1.3.3. Elaborarea planului de sondaj 50
1.4. PROBLEME ORGANIZATORICE 57
1.4.1. Asigurarea necesarului uman 57
1.4.2. Asigurarea mijloacelor materiale şi financiare. Bugetul anchetei 59
1.4.3. Calendarul cercetării 62
1.4.4. Organizarea spaţiului 63
1.5. EXERCIŢII 63

2. SONDAJUL ALEATOR SIMPLU 67
2.1. NOŢIUNI ŞI NOTAŢII 67
2.1.1. Populaţie şi eşantion. Parametri şi estimaţii 68
2.1.2. Procedee de selecţie 72
2.1.3. Selecţii aleatoare. Estimatori. Distribuţii de selecţie 77
2.2. STATISTICI UTILIZATE ÎN CAZUL SONDAJULUI ALEATOR SIMPLU 83
2.2.1. Media de selecţie 83

2.2.2. Valoarea globală sau totalul de selecție 89
2.2.3. Moment inițial de selecție de ordin k 90
2.2.4. Proporția de selecție 90
2.2.5. Varianța de selecție 92
2.2.6. Moment centrat de selecție de ordin k 94
2.3. SELECȚII PENTRU POPULAȚII CARE URMEAZĂ LEGI DE DISTRIBUȚIE NORMALE 94
2.4. ESTIMAREA PRIN INTERVAL DE ÎNCREDERE A UNUI PARAMETRU 98
2.4.1. Estimarea mediei 99
2.4.2. Estimarea valorii globale (a totalului) 106
2.4.3. Estimarea proporției 108
2.4.4. Estimarea variației 111
2.5. CALCULUL VOLUMULUI EȘANTIONULUI 113
2.5.1. Cazul unei variabile alternative 114
2.5.2. Cazul unei variabile numerice continue 117
2.6. EXERCIȚII 118

3. SONDAJUL NON-ALEATOR (EMPIRIC) 123
3.1. PRINCIPIILE SELECȚIEI NON-ALEATOARE 124
3.2. SONDAJUL PE COTE 126
3.2.1. Principiul de bază. Planul de sondaj 126
3.2.2. Probleme organizatorice specifice 132
3.2.3. Avantaje și limite 134
3.2.4. Ameliorarea calității unui sondaj pe cote 135
3.3. METODA UNITĂȚILOR TIP 136
3.4. ALTE METODE 137
3.4.1. Metoda voluntariatului 137
3.4.2. Metoda selecției bazate pe accesibilitate 138
3.4.3. Metoda itinerariilor 139
3.5. EXERCIȚII 139

4. SONDAJE MIXTE 143
4.1. JUSTIFICAREA METODELOR MIXTE 143
4.2. SONDAJUL STRATIFICAT 145
4.2.1. Principiul metodei 145
4.2.2. Estimarea parametrului medie 149
4.2.3. Construirea eșantionului 158
4.3. SONDAJUL ÎN DOUĂ SAU MAI MULTE TREPTE 164
4.3.1. Fundamentarea metodei 164

4.3.2. Proprietățile estimatorilor. Precizia rezultatelor 166
4.4. SONDAJUL PE SERII (CLUSTER) 175
4.4.1. Justificarea metodei 175
4.4.2. Calitatea estimatorilor și a rezultatelor 176
4.5. EXERCIȚII 180

5. UTILIZAREA UNEI INFORMAȚII SUPLIMENTARE. AMELIORAREA CALITĂȚII ESTIMATORILOR 183
5.1. PRINCIPIUL AMELIORĂRII ESTIMATORILOR PRIN REDUCEREA VARIANȚEI 183
5.2. POST-STRATIFICAREA 185
5.2.1. Proprietățile estimatorului 186
5.2.2. Limitele și eficiența metodei 189
5.3. ESTIMATORUL RAPORT (RATIO) 192
5.3.1. Principiu 192
5.3.2. Proprietățile estimatorului 195
5.4. ESTIMATORUL REGRESIE 199
5.4.1. Principiu 199
5.4.2. Proprietățile estimatorului 200
5.5. EXERCIȚII 204

TABELE PROBABILISTE 207

BIBLIOGRAFIE 219

Introducere

Our knowledge, our attitudes, and our actions are based to a very large extent on samples. This is equally true in everyday life and in scientific research.

W.G. Cochran

Ca și în ediția anterioară, în această lucrare prezentăm câteva elemente de bază ale teoriei și ale practicii sondajului statistic, care pot fi utile cititorului specializat sau celui ce dorește să se inițieze în acest domeniu. O nouă ediție a acestei cărți aduce câteva modificări față de prima versiune. Sunt aduse completări atât cu privire la elementele teoretice și metodologice, cât și cu privire la exemple și exerciții. Structura cărții nu se modifică față de prima ediție, însă au fost introduse noi elemente care să faciliteze demersul însușirii cunoștințelor privind cercetarea pe bază de sondaj statistic.

Subliniem remarca lui W.G. Cochran potrivit căreia societatea actuală este marcată în mod semnificativ de practica sondajelor statistice. Aproape fără excepție, metoda cercetării prin sondaj este folosită în toate științele și în activitatea socio-economică obișnuită a instituțiilor societății moderne. Argumentele unei asemenea prezențe sunt legate îndeosebi de puterea de cunoaștere pe care o oferă sondajul statistic. Există însă și situații în care sondajul reprezintă singura posibilitate de a realiza o cercetare științifică asupra realității (sunt cunoscute exemplele din medicină și industria farmaceutică, din domeniul controlului calității produselor sau cele cu privire la opinia populației etc.).

În literatura de specialitate se întâlnesc lucrări ce prezintă sondajul statistic din perspective foarte diverse. Există lucrări, scrise din perspectiva statisticianului, care sunt foarte tehnice și din care lipsesc elemente de metodologie privitoare la partea de elaborare a unei anchete prin sondaj sau elementele organizatorice specifice acestei cercetări. De cealaltă parte, se întâlnesc situații în care sondajul este prezentat de o manieră simplificată, fără prea multe detalii legate de fundamentarea statistică a tehnicilor de sondaj. Asemenea lucrări sunt scrise din perspectiva cercetătorilor din domeniul științelor sociale, interesul acestora fiind legat de utilizarea metodelor de sondaj pentru diverse obiective de cercetare.

Intenția acestei lucrări este aceea de a oferi cititorilor o carte introductivă cu privire la teoria și practica sondajului statistic, care ia în considerare observațiile de mai sus. Cartea se adresează tuturor cercetătorilor, indiferent de nivelul de cunoștințe de statistică și matematică. În primul capitol al lucrării sunt prezentate elemente metodologice și organizatorice strict necesare pentru a putea elabora o cercetare prin sondaj. Accentul este pus pe etapa de elaborare a temei de cercetare, a instrumentului de culegere a datelor și a unui plan de sondaj. În capitolele următoare sunt prezentate câteva metode de sondaj împreună cu fundamentarea statistică, fără de care nu se poate garanta condiția de reprezentativitate a eșantionului și, prin urmare, calitatea rezultatelor. Astfel, în capitolul al doilea se discută detaliat cea mai cunoscută metodă de sondaj statistic: sondajul aleator simplu. În al treilea capitol sunt prezentate câteva metode de sondaj empirice sau non-aleatoare. Urmează un capitol dedicat planurilor de sondaj mixte care combină principiile de selecție ale metodelor prezentate anterior, adică aleator și empiric. Lucrarea se încheie cu prezentarea unor metode de îmbunătățire a calității rezultatelor sondajelor prin utilizarea unei informații suplimentare în etapa de estimare a parametrilor.

Iași, iulie 2018 *Dănuț-Vasile Jemna*

Capitolul 1

FUNDAMENTELE CERCETĂRII PRIN SONDAJ

Ca principiu de cunoaștere, sondajul este utilizat în toate domeniile vieții sociale și economice, atât în forme intuitive, neștiințifice, cât și în contextul unor cercetări ce presupun rigoarea metodologică a științelor moderne. Ca metodă de cercetare științifică, sondajul face parte din categoria cercetărilor statistice, cantitative și este utilizat în toate domeniile cunoașterii: economie, științe sociale, științe ale naturii, medicină. Sondajul s-a impus ca o posibilitate de cunoaștere a unei *populații* statistice (de regulă, definite ca un ansamblu de unități reale de volum mare) prin studierea unei părți a acesteia, numite *eșantion*, extrase din ansamblu după un procedeu care asigură condiția de *reprezentativitate*. Rezultatele obținute prin analiză statistică la nivelul eșantionului sunt apoi *extrapolate* la nivelul întregului, pentru a atinge obiective precum estimarea unor parametri sau verificarea unor ipoteze statistice formulate cu privire la populația analizată.

În acest capitol ne propunem o introducere cu privire la fundamentele cercetării prin sondaj. Ne interesează câteva coordonate istorice, prin care se subliniază condițiile care au favorizat apariția și dezvoltarea acestei metode de cercetare. Secțiunea următoare este dedicată conceptului de reprezentativitate a unui sondaj. Partea cea mai elaborată a acestui capitol vizează prezentarea unor probleme teoretice și metodologice ale sondajului, potrivit demersului său metodologic: formularea unei probleme de cercetare, construirea chestionarului statistic și elaborarea unui plan de sondaj. În ultima secțiune sunt prezentate câteva aspecte organizatorice, care corespund punerii în practică a cercetării prin sondaj.

1.1. De la cercetările exhaustive la cele selective

Viața de zi cu zi este marcată de statistică și de studiile pe bază de sondaj. În permanență, instituții guvernamentale și firme private organizează sondaje statistice pentru a culege date din industrie, agricultură, comerț, servicii, medicină, demografie, sănătate, politică, divertisment, cultură și societate etc. Cercetarea pe bază de sondaj statistic, deși atât de cunoscută și larg folosită în prezent în toate științele, are o istorie recentă. Secolul al XX-lea poate fi numit secolul sondajelor, pentru că în această perioadă s-au dezvoltat fundamentele teoretice și metodologice ale acestei metode de cercetare. Publicul larg s-a obișnuit deja cu terminologia specifică metodei, pentru că rezultatele anchetelor prin sondaj sunt în permanență prezentate opiniei publice prin mass-media și studii de specialitate.

1.1.1. Terminologie

În societatea actuală, termenii de specialitate din diverse științe sunt popularizați prin mijloacele de comunicare în masă. Acest aspect este valabil mai ales cu privire la statistică. Publicul larg este familiarizat cu noțiuni specifice cercetării pe bază de sondaj pentru că o cantitate importantă de informații este obținută cu ajutorul acestei metode de investigare a realității și apoi este diseminată în mass-media. Terminologia legată de sondajul statistic este prezentă în spațiul public nu numai în limba română, ci și în limbile de circulație internațională. În bagajul culturii generale se regăsesc termeni precum eșantion, exit-poll, eroare, estimație, sondaj de opinie, probabilitate etc.

Înainte de a ne fixa asupra unei definiții științifice, putem sublinia că termenul *sondaj* este prezent în vocabularul comun cu diverse nuanțe și este utilizat cu sensuri specifice în diverse domenii ale activității umane. Dacă pentru statistică s-a impus un termen generic în toate limbile moderne de circulație internațională (*statistics, statistique, statistik* etc.), pentru sondaj lucrurile nu sunt la fel. Ca etimologie, termenul provine din limba franceză (*sondage*), având sensul de a sonda sau a analiza un mediu oarecare (*Larousse*, 1995, p. 948). Originile termenului sunt legate de substantivul *sonde*, care provine de la un termen nordic ce desemnează un instrument de măsurare a adâncimii apei. Sensului maritim al termenului i se adaugă apoi, succesiv, sensul medical (sonda chirurgicală), sensul gnoseologic (a *sonda*, a încerca să afli gândurile, sentimentele, intențiile cuiva, a iscodi, a tatona, conform *DEX*, București, p. 1003), sensul geologic, sociologic etc.

Sensul modern și științific al termenului *sondaj* este legat îndeosebi de cercetarea statistică. În acest context, sondajul desemnează o metodă de cercetare parțială a unei colectivități. Vorbim despre o acțiune prin care se analizează un eșantion extras dintr-o populație, iar rezultatele obținute sunt extinse la nivelul

întregului. Cu acest sens, termenul se impune atât în literatura de specialitate, cât şi în limbajul comun, în cadrul limbilor francofone, graţie aplicării pe scară largă a anchetelor de opinie publică, îndeosebi după anul 1930.

În literatura de limbă engleză există o altă terminologie, care are la bază termenul *sample*, corespondentul pentru *eşantion*. De altfel, în această limbă s-au consacrat termeni diferiţi pentru activităţi din domenii distincte, dar care utilizează acelaşi principiu. De exemplu, în sociologie, pentru sondajul de opinie avem echivalentul englez *poll*, iar, în ştiinţele economice, pentru sondajul statistic avem termenul *sampling*.

1.1.2. Scurt istoric şi condiţii pentru dezvoltarea cercetării prin sondaj

Practica de a „sonda", adică de a face o evaluare pe seama unei mostre, a unei părţi dintr-un întreg, este una foarte veche şi, în acelaşi timp, una foarte actuală (de la banala degustare până la analiza vizuală a unei lucrări artizanale sau a recoltei de pe un ogor etc.). Această analiză corespunde unei evaluări intuitive asupra unor caracteristici ale realităţii, având convingerea că informaţiile obţinute pentru partea analizată sunt valabile şi pentru întreg. Sondarea realităţii prin apelul la părţi ale acesteia are de a face cu principiul eficienţei. Nu putem cunoaşte realitatea în mod exhaustiv decât în anumite cazuri şi nu dispunem decât de resurse financiare, de timp, umane etc. limitate.

Încercări de estimare a unor mărimi prin observarea unei părţi din realitate sunt întâlnite chiar din antichitate (un exemplu îl oferă Herodot în *Istoria* sa, prezentând cum anume regele Persiei estimează numărul soldaţilor în campania de cucerire a Greciei). Primele tentative de extrapolare a rezultatelor parţiale bazate pe raţionamente şi metode ştiinţifice datează din secolul al XIV-lea. În această perioadă de timp au loc diverse încercări de estimare a numărului populaţiei pornind de la conceptul de „foc", care face referire la o locuinţă, familie sau gospodărie (Droesbeke et al., 1987, p. 4). Din această perioadă datează primele înregistrări ale fenomenelor demografice în acte de stare civilă, activitate care asigură principala sursă de date statistice. Rezultatele cele mai semnificative, însă, apar în secolul al XVII-lea, odată cu formarea *Şcolii Aritmeticii Politice Engleze*. Prin folosirea *tehnicii multiplicatorului*, W. Petty (1623-1687) şi mai apoi Laplace (1749-1827) ajung la rezultate spectaculoase în estimarea a numeroşi indicatori demografici şi economici (de exemplu, Petty estima populaţia Londrei, în 1686, calculând mai întâi o estimaţie de 88 de mii de case, care adăpostesc aproximativ 695 de mii de persoane). La început, având un caracter intuitiv şi logic, metoda de cercetare parţială capătă mai apoi o justificare teoretică prin apariţia şi dezvoltarea teoriei probabilităţilor. Metoda este astfel îmbunătăţită ulterior prin contribuţiile lui

Laplace și ale altor matematicieni și a fost utilizată cu succes în câteva cazuri în Franța și Anglia, în secolul al XVIII-lea. În ciuda acestor realizări, secolul următor avea să rămână în istorie ca o perioadă dominată de metoda recensământului. Noile state moderne europene, motivate de nevoia de cunoaștere exhaustivă a realității, au organizat regulat recensăminte prin intermediul instituțiilor de statistică înființate la nivelul administrației.

Formalizarea și utilizarea pe scară largă a metodelor de cercetare prin sondaj este una târzie (este legată de prima parte a secolului al XX-lea), dacă ne gândim că primele încercări de extrapolare și primele formalizări teoretice ale cercetării parțiale datează din secolul al XVII-lea. O explicație a acestui fapt, dintr-o perspectivă a istoriei și sociologiei științei, ar fi aceea că nu au existat condițiile cognitive (științifice și filosofice) și sociale (pragmatice și politice) care să facă posibilă această dezvoltare. Ne oprim în continuare asupra unor repere și condiții care au creat cadrul dezbaterilor asupra cercetării prin sondaj la sfârșitul secolului al XIX-lea și începutul secolului al XX-lea.

a. Interesul politic și administrativ

Deși secolul al XIX-lea a fost dominat de cercetările pe bază de recensământ, nevoia de cunoaștere a realității se diversifică, iar necesarul de informații devine tot mai mare și mai divers. Este clar că metoda exhaustivă nu mai poate ține pasul cu asemenea nevoi. Acest lucru devine tot mai clar și pentru administrație. După apariția statelor moderne europene, se conturează ideea de stat ca diversitate geografică, demografică, socială, economică din interiorul unei granițe. A cunoaște statul înseamnă a lua în considerare teritoriul și diversitatea sa. Această idee va fi valorificată din plin de autori precum statisticianul N.A. Kiaer (1895), care vorbește despre reprezentativitate în sens teritorial. O investigație reprezentativă înseamnă, la Kiaer, o explorare a unui număr de localități, distribuite, pe întreg teritoriul țării, în așa manieră încât subansamblul localităților cercetate să fie o miniatură a întregului teritoriu. La nivel administrativ, prin instituțiile de statistică oficială, interesul statului cu privire la diverse probleme care privesc guvernarea a condus la dezvoltarea metodelor selective de culegere și analiză a datelor. Asemenea metode s-au dovedit a fi apoi nu numai singurele eficiente sub raportul timpului și costului, ci și singurele posibile.

Diversificarea nevoii de cunoaștere, împreună cu posibilitățile și limitele ce însoțesc activitatea administrativă, a condus la utilizarea pe scară largă a sondajului la nivelul instituțiilor oficiale de statistică. Actualmente, cea mai mare parte a datelor necesare guvernării se obțin cu ajutorul sondajelor statistice oficiale. Acestea au caracter reprezentativ la nivel național și se organizează cu regularitate după metodologii științifice riguroase. Rezultatele anchetelor sunt puse atât la

dispoziția guvernelor, cât și a instituțiilor de învățământ și cercetare și chiar a publicului larg (sub formă de indicatori sintetici prezentați în diverse publicații de popularizare).

b. Condițiile sociale și economice

Noul stat democratic deschis de modernitate a dus la dezvoltarea unor instrumente juridice și politice importante, cum ar fi codul civil și sufragiul universal. Mass-media devine o putere în statul modern, realizând posibilitatea exprimării a ceea ce numim opinie publică. Consultarea opiniei populației în diverse probleme, mai ales a opiniei politice, reprezintă o problemă actuală de democrație care a dus la dezvoltarea anchetelor de opinie și a campaniilor electorale naționale prin mass-media. America este țara care a dat tonul și modelul unei astfel de societăți încă din secolul al XIX-lea. În acest secol, în SUA, se realizează primele sondaje de opinie, inițiativă care vizează obținerea de informații despre opinia populației în legătură cu politicul. Motorul acestor anchete îl constituie organismele de presă, iar metoda este simplă: consultarea opiniilor personale prin presă. Primele anchete debutează în 1824, prin jurnalele *Pennsylvenian Harrisburg* și *Raleigh Star*, dar fără să pună problema reprezentativității (Droesbeke et al., 1987, p. 14). În istorie însă au intrat anchetele din anul 1936, realizate de diverse jurnale americane în preajma alegerilor prezidențiale. Două rezultate au atras atenția: *Literary Digest*, care, realizând un sondaj de presă pe un eșantion de peste 2 mil. de persoane, a prezis câștigarea alegerilor de către candidatul A. Landon, și institutul de sondaje al lui G. Gallup, care, după o metodă de sondaj rațional, pe un eșantion mic, a prezis câștigarea alegerilor de către Roosevelt. Deși sondajul realizat de Literary Digest a avut la bază un eșantion de volum mare, rezultatele au fost departe de adevăr, eșantionul fiind nereprezentativ (erori de sondaj semnificative). Practica anchetelor realizate de diverse jurnale s-a extins din Statele Unite ale Americii în Europa, la începutul secolului al XX-lea, devenind o modalitate de consultare a opiniei populației privind viața politică și socială. În paralel cu activitatea ziarelor, cercetarea pe bază de sondaj prinde contur la nivelul unor instituții specializate sau al unor companii private care au ca obiectiv sondarea opiniei publice.

Criza economică din Anglia, la sfârșitul secolului al XIX-lea, a determinat oficialitățile să studieze îndeaproape fenomenul sărăciei. Studiile engleze asupra sărăciei (Desrosieres, 2000, p. 272) și asupra problemelor sociale ridicate de industrializare și de concentrarea în orașe au adus contribuții majore privind clarificarea metodei statistice de cercetare prin sondaj. Pe lângă alte aspecte, aceste studii au pus problema măsurării erorilor care apar în anchetele statistice, precum și

a utilizării teoriei probabilităților în rezolvarea unor asemenea probleme metodologice.

Dezvoltarea pieţelor naţionale, datorată revoluţiei industriale şi dezvoltării infrastructurii, a condus la o nevoie de cunoaştere a proceselor din cadrul vieţii economice. Această nevoie trimite la dezvoltarea metodelor de cercetare a pieţei. În acelaşi timp, posibilitatea de a cunoaşte rezultatele de pe piaţă prin sondaje impune o standardizare a produselor pe piaţa naţională. Acest mecanism va continua să se dezvolte permanent: realitatea socio-economică ridică mereu probleme care stimulează ştiinţa să aducă soluţii în privinţa dezvoltării celor mai eficiente modalităţi de cunoaştere a acesteia; în egală măsură, nevoia de cunoaştere a complexităţii realităţii a condus la soluţii practice în plan economic şi social care să ţină cont de această nevoie şi să faciliteze acest proces. Începând cu prima parte a secolului al XX-lea se dezvoltă primele firme specializate în realizarea de studii de marketing. Între pionierii care au dus la dezvoltarea conceptului de cercetare de piaţă pe bază de sondaj se află şi George Gallup, cel care a pus bazele celebrei companii ce îi poartă numele.

1.1.3. Opţiunea între sondaj şi cercetarea exhaustivă

Este larg acceptat faptul că sondajul statistic şi-a câştigat locul aproape privilegiat în rândul metodelor de cercetare cantitativă atât în cadrul instituţiei oficiale de statistică, cât şi în universităţi şi instituţii de cercetare, în companii şi firme de producţie şi servicii. Analizând procesul prin care sondajul s-a impus ca o metodă ştiinţifică de cercetare de succes, putem delimita o serie de criterii care susţin opţiunea pentru metoda selectivă în raport cu cea exhaustivă. Prezentăm mai jos câteva dintre aceste criterii.

a. Tema de cercetare

În comparaţie cu cercetarea exhaustivă, prin sondaje se pot studia cu succes teme dintre cele mai diverse şi din domenii unde studiul întregii populaţii nu se poate realiza decât în condiţii limitate sau este chiar imposibil de aplicat. Există subiecte sau teme de cercetare care nu pot fi studiate decât prin apelul la sondajul statistic. Un exemplu îl reprezintă opinia populaţiei cu privire la opţiunea politică sau cu privire la o problemă socială sau economică. De asemenea, comportamentul şi consumul populaţiei în contextul economiilor de piaţă reprezintă un alt exemplu. Este destul de clar că nu pot fi intervievate toate persoanele dintr-o anumită ţară, regiune sau chiar localitate pentru a obţine informaţiile dorite pentru asemenea subiecte de mare interes. Un alt exemplu îl constituie studiile privind calitatea produselor obţinute într-o companie (aici un caz limită îl reprezintă situaţia în care cercetarea presupune chiar scoaterea din uz a produselor verificate).

b. Obiectivele propuse

În cercetarea prin sondaj se pot realiza studii analitice ale fenomenelor, cu posibilitatea stabilirii unor legi, modele şi previziuni, ceea ce nu e cazul unei cercetări exhaustive, care rămâne, cel mai adesea, la stadiul descriptiv. Dincolo de limitele impuse de cercetarea cantitativă, sondajele statistice oferă oportunitatea de a evalua empiric teoriile formulate de specialiştii din diverse domenii ale cunoaşterii, de a explora noi probleme şi de a formula noi ipoteze, de a asigura suportul pentru luarea deciziilor etc.

c. Costul şi timpul de realizare a cercetării

În cercetările statistice, o constrângere importantă o reprezintă costul. Deşi costul cercetării pe unitate statistică este superior într-o cercetare prin sondaj, pe ansamblu, costul unui sondaj este mult inferior costului unei cercetări exhaustive. Realizarea periodică a studiilor prin sondaj conduce însă la o scădere a costului unitar de până la două sau trei ori, pentru că se utilizează logistica şi informaţiile anterioare. În această situaţie, unele anchete prin sondaj se pot realiza într-un timp foarte scurt, ceea ce reprezintă un aport substanţial al sondajului în condiţiile actuale de piaţă. Apelul la anchetele permanente (de tip panel) este o soluţie extrem de productivă, atât în privinţa rapidităţii şi actualităţii datelor, precum şi a costului.

d. Precizia

Realizate pe un număr restrâns de cazuri, sondajele pot permite o rigoare înaltă cu privire la calitatea activităţilor desfăşurate. În realizarea sondajului, de la etapa de observare şi până la întocmirea raportului, pot lucra persoane specializate, cu competenţe bine definite. Ca rezultat, calitatea muncii şi precizia rezultatelor sondajului sunt superioare recensământului. În aceeaşi ordine de idei se înscriu şi: posibilitatea realizării de pre-anchete care să pregătească cercetarea prin sondaj; realizarea de post-anchete care să realizeze controlul rezultatelor; posibilitatea stimulării financiare sau materiale a persoanelor anchetate astfel încât să ofere date corecte; posibilitatea unui control riguros al datelor înregistrate; posibilitatea ameliorării calităţii rezultatelor etc.

1.1.4. Limite ale utilizării sondajelor

Ca metodă de cercetare ştiinţifică, sondajul admite o serie de limite şi necesită o serie de condiţii de aplicare specifice. Prezentăm succint câteva aspecte din această arie tematică.

a. Posibilitatea de a asigura *reprezentativitatea eşantionului* reprezintă piatra de încercare în organizarea anchetelor prin sondaj. Cercetările prin sondaj statistic nu pot fi aplicate decât în cazul în care această proprietate este îndeplinită. Această condiţie impune existenţa unor condiţii preliminare, cum ar fi: o bază de date cu

elemente de identificare pentru toate unitățile populației; informații despre structura și omogenitatea populației; informații din studii anterioare; criterii și caracteristici care asigură o corespondență clară între populație și eșantion etc. În practică, deseori este dificil de asigurat condiția de reprezentativitate a eșantionului, iar soluțiile la această problemă determină alegerea unei anumite metode de sondaj, a unui nivel de precizie a rezultatelor, o anumită perioadă de timp de realizare sau un anumit cost al cercetării. Pentru o populație reală, care, în general, este una complexă, erorile de reprezentativitate sunt dificil de evaluat, iar proprietatea de reprezentativitate nu este ușor de asigurat.

b. *Costurile*, dar mai ales *rigorile metodologice* cerute de anumite metode de sondaj (cum ar fi cazul sondajelor aleatoare) sunt uneori semnificative și nu pot fi îndeplinite în totalitate. Acest lucru determină opțiunea pentru alte tipuri de sondaj decât cele așteptate (sondaje empirice, de exemplu, ale căror rezultate nu pot fi apreciate prin măsurători statistice sub aspectul preciziei) sau decizia de a aplica alte metode de analiză, cum ar fi cele calitative (de exemplu, interviuri).

c. Cu toate facilitățile unui sondaj, *analiza în profunzime* a unui fenomen nu este posibilă decât prin intermediul monografiilor (ca studii multidisciplinare și exhaustive a realității) sau a cercetărilor calitative. Sondajul permite o analiză a temelor supuse analizei în orizontul posibilităților oferite de abordarea cantitativă, care presupune măsurarea și evaluarea numerică. La acest punct, literatura de specialitate ne arată că taberele sunt împărțite, existând opinii pro și contra, studii teoretice și metodologice cu privire la abordarea cantitativă, în special la limitele de cunoaștere ale acesteia.

d. Rezultatele sondajului depind în mare măsură de *calitatea răspunsurilor* persoanelor anchetate. Atât mediatizarea anchetei în ce privește justificarea caracterului științific și interesul practic ale sondajului, precum și asigurarea condițiilor de observare statistică sunt probleme delicate de care depinde calitatea rezultatelor. În măsura în care cercetările prin sondaj acordă o mare atenție etapei de pregătire a anchetei și apoi celei de culegere a datelor, rezultatele corespund realității. Dacă se neglijează calitatea procesului de culegere a datelor, atunci rezultatele pot suferi semnificativ. În aceeași ordine de idei, pentru a asigura o calitate ridicată a rezultatelor este nevoie de un personal specializat, bine pregătit, nu numai în pregătirea teoretică și metodologică a anchetei, ci și în pregătirea etapei de culegere a datelor.

e. Selecția efectivă a unităților din populație nu este ușor de realizat, indiferent de metoda de sondaj aplicată. În practică, cel mai frecvent trebuie gestionată problema *non-răspunsurilor*. Pentru sondajele aleatoare sau mixte, unitățile nu sunt interschimbabile, fapt ce impune o supradimensionare prealabilă a

eşantionului pentru a asigura la final numărul necesar de unităţi în eşantion. În cazul sondajelor empirice, unităţile pot fi înlocuite, iar procesul de selecţie presupune un efort mai mare şi cu mai multe riscuri de a realiza erori de reprezentativitate.

1.2. Conceptul de reprezentativitate a unui sondaj

În cercetările pe bază de sondaj statistic, asigurarea reprezentativităţii[1] eşantionului constituie cel mai important criteriu de fundamentare teoretică şi metodologică. În esenţă, conceptul de reprezentativitate statistică pune următoarea problemă: în ce condiţii partea sau eşantionul poate *(re)prezenta*, înlocui întregul sau populaţia totală din care a fost extras? La această problemă s-au dat diferite soluţii, conceptul de reprezentativitate suportând o evoluţie semantică în strânsă legătură cu dezvoltarea problemelor teoretice şi practice ale cercetării statistice.

Reprezentativitatea a fost definită în principal de pe două poziţii. O primă abordare este de tip empiric, intuitiv, conceptul fiind definit în acord cu *modul* de manifestare a realităţii ca întreg, pe baza informaţiilor disponibile despre acest întreg. Cea de-a doua abordare a conceptului este de tip teoretic, pe baza unui set de *principii* şi *legi* care permit definirea şi extragerea unui eşantion din populaţie. Această poziţie s-a dezvoltat graţie aparatului oferit de teoria probabilităţilor şi de statistica matematică. Prezentăm pe scurt câteva elemente mai semnificative pentru fiecare tip de abordare.

1.2.1. Elaborări empirice

Aşa cum am arătat în subcapitolul anterior, tentativele de formalizare a unui suport logic şi ştiinţific pentru cercetările pe bază de sondaj datează din perioada secolelor XIV-XVII. Aceste rezultate vor cunoaşte însă o dezvoltare de amploare şi pe baze ştiinţifice riguroase abia la începutul secolului al XX-lea, într-o perioadă determinată de factori politici, economici şi ştiinţifici favorabili. Până atunci, însă, cercetările prin sondaj se realizează într-un cadru restrâns şi pe baze empirice, intuitive. Două modalităţi de definire a reprezentativităţii sau de justificare a cercetărilor prin sondaj reţin atenţia în perioada secolelor XVIII-XIX: reprezentativ înseamnă tipic şi reprezentativ înseamnă a păstra structura întregului la nivelul părţii sau al eşantionului.

---

[1] Se reia în acest subcapitol, cu anumite modificări, studiul „Conceptul de reprezentativitate în cercetarea statistică", publicat de autor în *Analele Univiversităţii „Al.I. Cuza" din Iaşi*, Tomul L/LI, 2004/2005, pp. 413-418.

a. Reprezentativ - tipic

În cea mai mare parte a anchetelor parţiale şi a monografiilor secolelor XVIII-XIX, reprezentativ înseamnă *tipic* (Desrosieres, 2000, p. 266). În această situaţie, reprezentativitatea este asigurată prin găsirea uneia sau a mai multor *unităţi tip* din populaţie. Unităţile tip reprezintă acele unităţi care au valorile variabilelor studiate foarte aproape de medie. Justificarea acestei abordări este una intuitivă. Dacă întregul este descris cu ajutorul mărimilor medii, atunci e suficient să se studieze acele unităţi din populaţie ale căror caracteristici se apropie de medie.

Unităţile tip sunt alese cu grijă, pe baza datelor obţinute în anchetele exhaustive. Din datele statistice oficiale se obţin informaţii cu privire la populaţia de referinţă, se obţin mărimile medii pentru caracteristicile de interes, iar pe baza acestor informaţii se identifică unităţile care vor fi supuse studiului monografic, unei analize în profunzime. Specificul anchetei monografice este calitatea, studiul complex, multilateral al unităţilor tip alese din populaţie.

Critica adusă acestui mod de a defini reprezentativitatea constă, în principal, în aceea că prin analiza unităţilor tip nu se ţine cont de diversitatea populaţiei, de structura acesteia după diferite criterii (Desrosieres, 2000, p. 279).

b. Reprezentativ - miniatură a întregului

Condiţia de reprezentativitate impune să se obţină un eşantion care să respecte structura populaţiei de referinţă şi, implicit, diversitatea existentă la nivelul acesteia. Când structura întregului este cunoscută, eşantionarea se poate realiza utilizând această informaţie, iar eşantionul apare ca o miniatură a populaţiei totale. Aceasta este viziunea despre reprezentativitate promovată de statisticianul norvegian N.A. Kiaer. La Congresul Institutului Internaţional de Statistică din 1895, de la Berna, Kiaer readuce în lumea statisticii internaţionale discuţia asupra problemei sondajului (Kiaer, 1895, pp. 176-178). Autorul pune problema reprezentativităţii astfel: cum să obţii un număr de unităţi dintr-un întreg astfel încât acesta să (re)prezinte în miniatură, cât mai fidel, colectivitatea totală. Criticând abordarea conceptului în contextul anchetelor din secolele XVII-XIX, Kiaer consideră că o anchetă reprezentativă trebuie să ia în considerare nu doar cazurile tipice, ci varietatea care se găseşte într-un anumit spaţiu social, propunând ca în sondaj să se realizeze o miniatură a acelui spaţiu sau ţări. Organizarea de anchete reprezentative în Norvegia (mobilul practic) a fost motivată de interesul statului cu privire la politicile de protecţie socială şi cu privire la necesitatea creării de organisme în acest sens.

Prin Kiaer, se realizează o trecere de la statistica întregului, caracterizată prin medie (Quetelet, 1835), la statistica analizei distribuţiei indivizilor, o statistică

bazată pe varianţă, corelaţie, comparaţie. Kiaer afirmă că totalul nu este dat de cazul tip, de medie, ci de toată varietatea cazurilor care se întâlnesc în viaţă. Însă abordarea lui Kiaer este mai mult intuitivă şi empirică. El este un bun cunoscător al muncii de teren, al realizării practice a anchetelor, fiind mai puţin preocupat de formalizarea matematică şi teoretică a metodei sale. În ciuda acestei lipse de preocupări pentru formalizare, discuţiile deschise de Kiaer sunt fundamentale, de pionierat. Kiaer nu a pus problema modului de extragere (aleatoare sau de alt tip) a unui eşantion, ci a sugerat modul în care se poate verifica reprezentativitatea, după metoda variabilelor de control. Rezultatele obţinute pentru aceste variabile sunt comparate cu cele obţinute prin recensământ. Dacă abaterile nu sunt semnificative, atunci rezultatele anchetei prin sondaj sunt considerate valide.

Puterea cognitivă a cercetării prin sondaj propusă de Kiaer se poate evidenţia în comparaţie cu metodele clasice: recensământul şi monografia. Faţă de recensământ, sondajul se impune prin două elemente specifice: posibilităţile de cunoaştere şi precizia. Dacă prin recensământ realitatea este analizată la suprafaţă, metoda reprezentativă permite un studiu în profunzime al realităţii. În primul caz, informaţiile obţinute sunt descriptive, pe când, în al doilea caz, acestea permit explicarea realităţii supuse cercetării. În plus, metoda reprezentativă este mai precisă, rezultatele fiind afectate de mai puţine erori. Atât sondajul, cât şi monografia analizează în profunzime realitatea, însă monografia se rezumă la cazurile tipice, pe când sondajul studiază întreaga diversitate a populaţiei.

Valoarea ştiinţifică a cercetării prin sondaj depinde de caracterul reprezentativ al eşantionului şi mai puţin de volumul acestuia. Deseori se întâmplă ca datele disponibile sau practic uşor de obţinut să nu se refere decât la o parte a populaţiei. A realiza o cercetare doar pe baza acestor date înseamnă a comite o eroare de reprezentativitate. Or, o eroare de acest tip au comis americanii în anchetele de opinie prin mass-media, mizând pe un volum mare de date, dar fără a cunoaşte gradul lor de reprezentativitate.

1.2.2. Elaborări ştiinţifice

Cu toate realizările lui Kiaer, metoda reprezentativă va pătrunde cu greu atât în teoria, cât şi în practica statistică. O primă cauză este de ordin teoretic: metoda lui Kiaer rămâne una intuitivă şi empirică, nu are un fundament ştiinţific, matematic. Deschiderea se realizează câţiva ani mai târziu prin utilizarea distribuţiilor de probabilitate la extragerea unui eşantion şi la compararea rezultatelor de sondaj cu cele de la nivelul întregii populaţii. O altă cauză a demarajului greoi al metodei sondajului statistic este dată de reţinerea comunităţii

ştiinţifice cu privire la sondaj, sub impactul şi influenţa administraţiei şi statisticii oficiale, bine înrădăcinate în tradiţia recensămintelor.

După anul 1925, când are loc Congresul Institutului Internaţional de Statistică de la Roma, se apreciază că *metoda reprezentativă* este validă, iar problemele care se deschid ţin de modalităţile de obţinere a unui eşantion din populaţia totală. Pe baza rezultatelor obţinute de specialişti în prima parte a secolului al XX-lea, conceptul de reprezentativitate a fost definit, în sens restrâns, în legătură cu procedeele aleatoare sau mixte de extragere a unui eşantion. În acest cadru, se remarcă următoarele corelaţii conceptuale: reprezentativ-aleator, reprezentativ-precis, reprezentativ-stratificat.

a. Reprezentativ-aleator

În plan teoretic, abordarea relaţiei eşantion-populaţie cu ajutorul instrumentelor oferite de teoria selecţiei determină o altă accepţiune asupra reprezentativităţii. În acest caz, reprezentativ înseamnă să extragi eşantionul după modelul unei experienţe aleatoare, după o schemă probabilistă, caz în care fiecare unitate din populaţie are o probabilitate cunoscută de a intra în eşantion. Reprezentativitatea este garantată de principiul extragerii aleatoare, adică după un principiu care asigură calculul probabilităţii de includere a unei unităţi în eşantion. De exemplu, sondajul aleator simplu asigură fiecărei unităţi din populaţie aceeaşi şansă de a intra în eşantion, ceea ce înseamnă că eşantionul extras va *reprezenta* populaţia. Schema de extragere probabilistă permite, de asemenea, calculul unui volum al eşantionului care nu numai că este reprezentativ, ci asigură şi un anumit grad de precizie aşteptat.

b. Reprezentativ-precis

Dezvoltările din teoria probabilităţilor şi statistica matematică au permis determinarea unui tip de erori care însoţesc cercetarea prin sondaj statistic, şi anume erorile de reprezentativitate. Aceste erori reprezintă măsuri care iau în calcul diferenţele dintre valorile de sondaj (estimaţii) calculate la nivelul unui eşantion şi parametrii populaţiei din care a fost extras eşantionul. Cu cât erorile de reprezentativitate sunt mai mari, cu atât este afectată reprezentativitatea rezultatelor. În acest caz, reprezentativitatea statistică înseamnă obţinerea celor mai bune estimaţii sau rezultatelor cu erorile cele mai mici. Un eşantion este considerat reprezentativ dacă pentru o variabilă dată estimează valoarea adevărată sau parametrul cu o eroare minimă, adică cu eroare zero (Ardilly, 1994, p. 388).

c. Reprezentativ-stratificat

Fundamentarea reprezentativităţii statistice a suportat o nouă clarificare prin combinarea celor două posibilităţi de a obţine un eşantion: cea empirică şi cea

probabilistă. Utilizând informaţiile despre existenţa unor categorii diferite de unităţi în populaţia totală, pe de o parte, şi fundamentul matematic al extragerii aleatoare, pe de altă parte, J. Neyman elaborează metoda sondajului stratificat. Autorul arată superioritatea unui sondaj aleator faţă de unul raţionat, nealeator, precum şi posibilitatea de a ameliora rezultatele sondajului aleator printr-o stratificare a priori a populaţiei totale (Neyman, 1934, pp. 558-606). Obţinerea reprezentativităţii în cazul unui sondaj stratificat reclamă două exigenţe:

- construirea de nomenclatoare, clase care să asigure o stratificare a priori a populaţiei totale după variabile care asigură diferenţe semnificative între indivizii din clase diferite şi diferenţe cât mai mici între indivizii din aceeaşi clasă;
- construirea schemelor de extragere aleatoare din aceste clase a indivizilor şi definirea instrumentelor care să verifice gradul de reprezentativitate a rezultatelor obţinute (instrumente specifice statisticii inferenţiale).

În cazul sondajului stratificat, există mai multe posibilităţi de construire a eşantionului: din fiecare strat să se aloce acelaşi număr de unităţi, un număr proporţional de unităţi cu ponderea stratului din total sau un număr de unităţi care să ţină cont nu numai de pondere, ci şi de gradul de omogenitate al fiecărui strat (alocare optimală). Eşantionul obţinut prin stratificare, cu o alocare proporţională a unităţilor din fiecare strat, este numit *eşantion reprezentativ*, iar alocarea optimală mai poartă şi numele de alocare Neyman.

Concluzie

Conceptul de eşantion reprezentativ se referă la proprietatea fundamentală a unui eşantion de a permite descrierea şi explicarea populaţiei din care a fost extras în condiţii de calitate cunoscute. Aceasta este condiţia de bază a valabilităţii ştiinţifice a unei colectivităţi de selecţie (Bărbat, 1971, p. 436). Celor două modalităţi de abordare a conceptului de reprezentativitate le corespund două metode de cercetare prin sondaj distincte: metode raţionate sau empirice (sondajul pe cote, metoda unităţilor tip etc.) şi metode aleatoare. Prin combinarea celor două metode fundamentale, a rezultat un tip de sondaj mixt, care utilizează avantajele ambelor metode de sondaj (cazul sondajului stratificat, de exemplu).

Definirea reprezentativităţii unui eşantion se poate face atât în sens restrâns, cât şi în sens larg:

- în sens restrâns, un eşantion este reprezentativ dacă fiecare unitate din populaţie are posibilitatea de a intra în eşantion cu o probabilitate cunoscută. Cu alte cuvinte, este reprezentativ eşantionul care este extras după o metodă aleatoare. Definiţia se extinde şi la eşantioanele mixte, care implică o dimensiune aleatoare în procesul de selecţie;

- în sens larg, un eşantion este reprezentativ dacă permite descrierea şi explicarea populaţiei de referinţă cu erori cât mai mici. Potrivit acestei accepţiuni, reprezentativitatea poate fi extinsă şi la eşantioane obţinute după metode de sondaj nealeatoare.

Problema reprezentativităţii eşantionului este esenţială într-un sondaj statistic, însă rămâne o problemă sensibilă şi dificil de rezolvat. În practică, de regulă, reprezentativitatea unui eşantion se defineşte în raport cu o singură variabilă, cu o variabilă fundamentală pentru tema de cercetare. Problema practică cea mai dificilă este că nimic nu garantează faptul că eşantionul este reprezentativ şi în raport cu o altă variabilă, mai ales în cazul în care prin sondaj se cercetează o problemă complexă. O soluţie, pentru asigurarea reprezentativităţii eşantionului, este să se organizeze cercetări pe teme bine conturate, iar variabila prin care se defineşte reprezentativitatea să fie puternic corelată cu variabilele care apar în studiu. În cazul problemelor complexe se recomandă descompunerea problemei pe componente care pot face obiectul unor cercetări separate. În fundamentarea reprezentativităţii unui eşantion, pe lângă problemele care ţin de metoda de sondaj, esenţiale sunt şi elementele care ţin de: definirea populaţiei de referinţă sau a întregului, definirea unităţilor individuale de selecţie şi de observare, definirea caracteristicilor populaţiei după care se realizează cercetarea etc.

1.3. Probleme teoretice şi metodologice

În literatura de specialitate, etapele unei cercetări prin sondaj sunt prezentate în diverse forme, mai sintetic sau mai analitic. Lucrările clasice de sondaje prezintă câteva etape preliminare celei de culegere a datelor, cum ar fi: definirea obiectivelor cercetării, a populaţiei şi a datelor ce trebuie culese, identificarea informaţiei deja existente etc. (Cochran, 1997, p. 5; Droesbeke, 1997, p. 18). În această lucrare preferăm o abordare care consideră ca primă etapă a cercetării *definirea problemei* sau *a temei de cercetare*, utilizând o analitică ce nu este specifică lucrărilor cu privire la sondajul statistic. Preferăm să analizăm detaliat această primă etapă utilizând logica propusă în lucrările de metodologia cercetării din ştiinţele sociale, care pun accent pe etapa de fundamentare a temei de cercetare.

Etapele unei cercetări statistice bazate pe sondaj sunt prezentate schematic în Figura 1.1. În această schemă, după prima etapă a cercetării, putem delimita două tipuri de probleme care trebuie rezolvate în contextul realizării unui sondaj. Pe de o parte, vorbim despre problemele de ordin metodologic, surprinse în etapa elaborării planului de sondaj. Pe de altă parte, este vorba despre o serie de probleme de ordin organizatoric ce privesc elementele de realizare efectivă a cercetării. Conceperea chestionarului statistic este considerată o etapă distinctă în unele abordări, deşi

această etapă poate fi asociată cu cea a formulării temei de cercetare. În mod logic, chestionarul se construiește după ce au fost definite toate instrumentele conceptuale și metodologice ce vor fi utilizate în anchetă, inclusiv obiectivele și ipotezele de cercetare. Cu privire la elaborarea chestionarului, utilizăm o abordare proprie care face apel la trei dimensiuni de analiză: arhitectural (tehnici de construcție și elemente de structură); de conținut (tipologia întrebărilor și probleme specifice de elaborare); de impact (forma, problemele legate de editare).

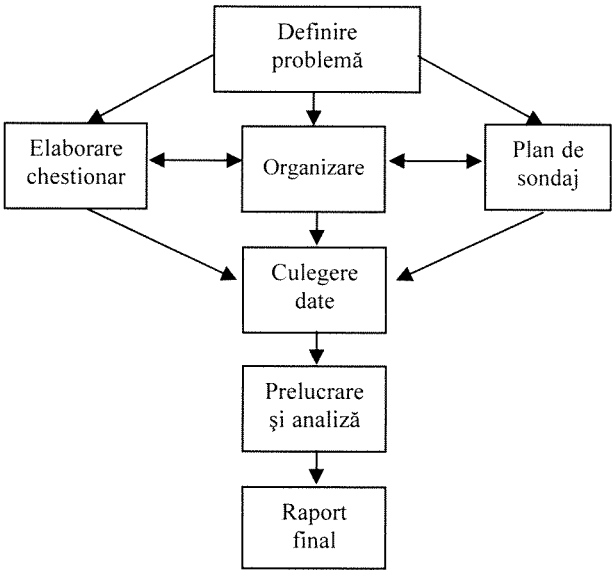


Figura 1.1. *Demersul metodologic al cercetării prin sondaj*

Asupra etapei de culegere a datelor nu insistăm, ci analizăm în treacăt câteva aspecte în prelungirea celei de construire a chestionarului. Ultimele două etape ale demersului metodologic sunt la fel de importante, însă nu vor fi abordate în lucrarea de față. De regulă, metodele de analiză a datelor sunt tratate în lucrări de statistică specifice, iar etapa de elaborare a raportului de cercetare se prezintă detaliat în lucrări de metodologia cercetării.

Dintre probleme teoretice și metodologice, în această lucrare tratăm pe scurt câteva aspecte privitoare la *definirea temei de cercetare*, *elaborarea chestionarului* și *elaborarea unui plan de sondaj*. În partea finală a acestui subcapitol analizăm câteva aspecte privind *organizarea* unui sondaj statistic, cu accent pe elaborarea unui buget și a unei diagrame privind încadrarea în timp a activităților ce presupun realizarea sondajului.

1.3.1. Definirea temei de cercetare

Tema de cercetare reprezintă, de regulă, un construct conceptual care apare la finalul unui proces care ar putea cuprinde mai multe etape: analiza cererii de sondaj, delimitarea fenomenului real, explorarea, conceptualizarea, construirea unui model de analiză (pentru diverse elemente care ar putea privi această etapă, a se vedea Quivi și Campenhoudt, 1995, p. 16; Mouton și Marais, 1990, p. 51). Prezentăm, pe rând, câteva aspecte pentru fiecare etapă în parte.

a. Analiza cererii de sondaj

Sondajele statistice sunt organizate ca urmare a unor nevoi reale de informații ce provin fie din partea administrației, fie din partea unor actori sociali și economici (companii private, instituții non-guvernamentale, instituții de învățământ și cercetare etc.). Asemenea nevoi de informații ce pot fi obținute prin anchete statistice se sistematizează în cereri de cercetare prin sondaj. Cererile formulate de destinatarii sondajelor cuprind o serie de obiective ce pot fi formulate în diverse variante, de la solicitări clar definite și precis formalizate până la obiective prezentate mai puțin precis, în termeni neștiințifici sau uneori neclari.

O primă imagine asupra temei de cercetare se obține prin analiza acestei cereri de sondaj pusă la dispoziție de instituția sau persoana care solicită efectuarea anchetei. În acest sens, este important să existe un dialog între echipa de cercetători și destinatarii studiului, pentru a identifica scopul principal al sondajului și modul în care vor fi concretizate rezultatele. Cerințele concrete formulate de clientul care comandă sondajul pot permite creionarea primelor coordonate ale temei de cercetare:

- nivelul științific al cercetării (descriptiv - care prezintă sintetic elementele esențiale ale realității observate, explorator - care poate propune noi ipoteze și teorii, evaluativ - care verifică ipoteze existente, explicativ - care identifică factori determinanți, legi și modele, predictiv - care oferă informații despre fenomenul analizat în contexte noi);
- persoanele care trebuie intervievate și eventual o serie de caracteristici de interes ale acestora;
- aria de investigație sau spațiul geografic care dimensionează problema;
- tipuri de date care trebuie culese în contextul anchetei;
- alte informații.

Exemplu

Considerăm un exemplu frecvent întâlnit, precum cel al situației unui candidat la primăria unui oraș A ce dorește să afle informații referitoare la receptarea activității sale în rândul cetățenilor și ce șanse are la următoarele alegeri.

În cererea de sondaj, destinatarul ar putea menționa că dorește să afle:

- gradul de cunoaștere a proiectului său electoral în rândul alegătorilor;
- care este opinia cetățenilor cu privire la rezultatele activității sale de primar în diverse domenii;
- care este sprijinul de care se bucură pentru un nou mandat;
- profilul alegătorilor care îl susțin și al celor care sunt împotriva sa;
- obiecțiile celor care au decis că nu îl votează;
- ce anume consideră cetățenii că reprezintă prioritățile administrative pentru viitorul mandat;
- care este profilul primarului dorit de comunitate;
- etc.

Analiza cererii de sondaj ne poate permite sublinierea concluziei că avem de a face, în esență, cu un studiu descriptiv care poate admite nuanțe explorative și, în final, se rezumă la extrapolarea unei informații cantitative de la nivelul eșantionului la nivelul populației de referință (estimarea unor parametri). Este destul de clar că persoanele care participă la anchetă sunt reprezentate de cetățenii localității *A* cu drept de vot (persoanele cu domiciliul stabil în acea localitate și cu vârsta de 18 ani și peste).

b. Delimitarea naturii fenomenului real

În demersul elaborării temei de cercetare, un aspect esențial îl reprezintă identificarea registrului de realitate la care face trimitere problema supusă analizei: un fenomen natural, social sau economic, o clasă de obiecte sau de procese, comportamentul unei categorii de persoane etc. În aceeași ordine de idei, potrivit metodei statisticii, fenomenele şi procesele reale analizate sunt observate ca întreguri structurale organizate sub formă de colectivităţi. Pe lângă acest specific al cercetării statistice, trebuie subliniat cel al cercetării prin sondaj. Culegerea datelor privind realitatea studiată se realizează de la nivelul unei populații umane care participă la un interviu și care trebuie definită în această etapă a cercetării. Uneori cele două populații coincid, adică persoanele intervievate sunt direct participante la realitatea supusă analizei, alteori unitățile intervievate sunt doar furnizoare de informații și au un rol pasiv în raport cu tema studiată.

În etapa de pregătire a sondajului, cu privire la fenomenul studiat, este necesar să se stabilească o serie de elemente care să asigure corespondența dintre realitatea studiată, populația statistică definită pe baza acesteia și populația ce participă efectiv la selecția și colectarea datelor:

- *natura* sau *esența populației statistice* corespunzătoare realității investigate. În practica cercetărilor statistice, vorbim despre colectivităţi de fapte, evenimente, opinii, obiecte, procese, fenomene, indivizi, grupuri, organizații etc. În fiecare caz, în funcție de tema studiată, se identifică o caracteristică

care definește natura populației și pe care o au toate unitățile (de exemplu, calitatea de consumator al unui produs sau cea de cetățean al unei țări);

- în paralel cu identificarea naturii populației statistice analizate, se definește *populația de respondenți* care participă la selecția și culegerea datelor. Din perspectiva culegerii datelor, unitățile individuale ale acestei populații pot fi simple (un angajat al unei companii) sau complexe (o companie), active (un cetățean cu drept de vot) sau pasive (un copil al unei familii pentru care raportează date unul dintre părinți). În sondajele statistice, populația statistică cu care se lucrează este, în final, reprezentată de populația unităților care sunt supuse procesului de culegere a datelor, adică populația respondenților;
- *modul de manifestare a realității analizate.* Cercetarea statistică se bazează pe observarea realității așa cum se manifestă concret, ceea ce presupune un efort de analiză a complexității problemei analizate, de detaliere a acesteia pe dimensiuni, componente, modalități de manifestare;
- *caracteristici ale populației supuse observării.* Se identifică o serie de caracteristici ale unităților populației respondenților care sunt relevante pentru tema de cercetare. Aceste caracteristici se grupează după natura lor, posibilitățile de observare sau de culegere a datelor etc;
- *durata și spațiul de manifestare* (o perioadă de timp și arealul geografic la nivelul căruia se desfășoară realitatea studiată). Există studii care se realizează ocazional, pentru un moment sau perioadă de timp (cross-sectional), dar există și studii care se reiau cu regularitate (longitudinale sau panel).

Exemplu

Cererea de sondaj, prezentată mai sus, indică relativ ușor natura realității studiate, anume *opinia* populației privind activitatea și profilul unui candidat la primăria unei localități. Potrivit dicționarului de sociologie, *opinia* este „un enunț care dă expresie opțiunii cognitive și afective a unei persoane, grup social sau colectivitate pentru un anumit punct de vedere cu privire la un fapt sau eveniment, o relație sau interacțiune socială" (Zamfir și Vlăsceanu, 1998, p. 402). În etapele ulterioare, se va realiza o analiză detaliată asupra conceptului de opinie și a posibilităților de operaționalizare pentru studiul propus în această anchetă.

Deoarece opiniile supuse analizei sunt exprimate de cetățenii români cu drept de vot, din orașul *A*, cercetarea prin sondaj propusă are ca populație de respondenți persoanele cu domiciliul stabil din localitatea de referință, cu vârsta de peste 18 ani și cetățenie română. Natura populației investigate este reprezentată de

această caracteristică de cetățean român cu drept de vot, ceea ce implică o condiție de vârstă (peste 18 ani).

Pentru tema de cercetare, sunt relevante o serie de caracteristici ale acestei populații care sunt numite caracteristici *socio-demografice*: vârsta, genul, profesia, ocupația, domeniul de activitate, nivelul de educație, venitul, starea civilă, religia etc.

De regulă, ancheta se realizează într-o perioadă de timp de câteva săptămâni, iar rezultatele vor avea acest referent temporal. De asemenea, fiind vorba despre un oraș, delimitarea spațială se realizează prin apelul la această caracteristică administrativ teritorială, identificată la nivelul unităților populației cu ajutorul domiciliului stabil ce figurează în actul de identitate.

c. Explorarea

Etapa exploratorie presupune efortul de plasare a temei de cercetare în contextul altor cercetări, al literaturii de specialitate care propune teorii și explicații asupra problemei studiate, dar și o conectare a cercetătorilor ce realizează sondajul, direct sau indirect, cu realitatea supusă analizei.

Această etapă presupune realizarea unor activități concrete de tipul: documentare, realizarea de interviuri exploratorii și apelul la alte metode de explorare a realității supuse cercetării.

Documentarea

Activitatea de documentare este esențială și presupune un efort de cercetare a unor documente publicate și accesibile cercetătorilor. Principalele documente și surse de documentare sunt:

- *studii și analize similare* anterioare, care sunt publicate sub forma unor articole de specialitate sau rapoarte de cercetare. De cele mai multe ori, o anchetă prin sondaj nu reprezintă o noutate absolută, ci beneficiază de experiența altora. Din aceste studii se pot desprinde anumite direcții de cercetare, ipoteze, concepte și probleme conexe. Unele informații sunt utile pentru definirea eșantionului (de exemplu, o estimație a unui parametru) și a instrumentului de culegere a datelor (chestionare utilizate în anchete similare), pentru stabilirea metodei de sondaj (ce metode au fost folosite anterior, ce informații asupra populației au fost utilizate, ce costuri, ce volum al eșantionului etc.), iar altele, pentru definirea cât mai bună a temei de cercetare, a obiectivelor și ipotezelor etc.;
- *lucrări de specialitate* publicate sub formă de cărți și monografii care tratează problema studiată sau probleme similare și studii diseminate în reviste cu vizibilitate națională și internațională. Aportul acestor lucrări este, în principal, de nivel teoretic și metodologic. Printr-o documentare

minuţioasă, se pot desprinde teorii şi abordări cu privire la tema studiată, precum şi metode şi tehnici utile în cercetare. Inventarul teoriilor, conceptelor, abordărilor se realizează structurat, pentru a putea fi folosite în etapa următoare;

- *documentele oficiale* ale diverselor instituţii ale statului sau instituţii non-guvernamentale. O serie de date şi informaţii privitoare la populaţia de referinţă pot fi obţinute din documente oficiale, baze de date disponibile etc. Un exemplu tipic la care se apelează constant îl constituie instituţiile de statistică (la nivel regional, naţional sau internaţional). Acestea oferă date statistice oficiale importante pentru desfăşurarea anchetelor prin sondaj;
- *alte documente* ce pot fi disponibile la nivelul instituţiilor publice sau al celei care solicită ancheta.

Ca tehnică de bază, documentarea presupune construirea unor fişe de lectură care sistematizează informaţia după mai multe criterii: concepte, teorii, definiţii de lucru, metode şi metodologii, abordări, obiective şi ipoteze etc.

Interviuri exploratorii

Aceste interviuri ajută la conturarea problematicii cercetării, asigură contactul direct cu fenomenul studiat, ajută la identificarea unor ipoteze de lucru şi la clarificarea unor aspecte privind problema studiată (Quivi şi Van Campenhoudt, 1995, p. 63). De regulă, pentru atingerea acestor obiective se aplică interviurile semi-structurate. În aceste interviuri se utilizează un ghid de interviu sub forma unui set minim de întrebări elaborate de cercetător, dar admite şi formularea unor întrebări noi, care se nasc în momentul dialogului.

Interviurile exploratorii vizează obţinerea de informaţii cu privire la tema de cercetare de la cel puţin trei categorii de persoane care au un rol special prin poziţia pe care o ocupă în raport cu tema. Mai întâi, sunt vizaţi *specialiştii* în domeniu, cei care au expertiza necesară cu privire la tema de cercetare (de la cercetători şi profesori până la specialişti care lucrează în companii, instituţii publice etc.). În al doilea rând, sunt intervievate persoane care sunt *implicate direct* în realitatea analizată şi care cunosc foarte bine fenomenul studiat (sunt vizate persoane cheie, care cunosc foarte bine fenomenul studiat sau au competenţe şi pot realiza o analiză pertinentă). În al treilea rând, sunt vizate persoane care fac parte din publicul direct *interesat* de rezultatele anchetei (de la persoane care fac parte din instituţia care a comandat sondajul, până la cele din spaţiul public - media, instituţii de învăţământ etc.).

După realizarea interviurilor, are loc analiza acestora cu ajutorul metodei analizei de conţinut. Informaţiile obţinute sunt organizate pentru a servi etapelor următoare ale cercetării.

Metode exploratorii complementare

Pentru a identifica noi aspecte cu privire la problema analizată, se pot aplica o serie de metode de explorare în plus (Quivi și Van Campenhoudt, 1995, p. 78):

- *observare directă*. Dacă tema o permite, membrii echipei de cercetare pot realiza o observare directă a realității studiate, o observare participativă. Cu această ocazie, sunt culese date și informații directe care sunt utilizate în etapele ulterioare;
- *consultarea unor documente speciale*. Există cazuri în care, pentru definirea riguroasă a temei de cercetare, este necesar să se apeleze la o serie de documente tehnice, detaliate sau la altele mai speciale care nu fac parte din rândul celor întâlnite în biblioteci sau instituții publice. Se pot încadra aici documente rare sau private care pot oferi informații particulare;
- *alte interviuri*. În funcție de nevoie, la interviurile exploratorii realizate deja se mai pot adăuga și altele, organizate punctual, cu persoane mai speciale sau cu subiecte bine delimitate. Cu această ocazie se pot verifica anumite informații culese sau se pot pune întrebări suplimentare.

Exemplu

Pentru ancheta propusă în acest capitol cu titlu didactic, etapa de documentare se realizează relativ ușor pentru că vorbim despre un sondaj cu o temă cunoscută și care se realizează frecvent. Conceptele de opinie publică, în general, și cel de opinie politică, în particular, sunt studiate în lucrări de sociologie și politologie, iar în aceste studii se pot identifica ușor dimensiunile conceptului, teorii explicative, factorii determinanți, instrumentele de măsurare etc.

O sursă importantă de documentare o reprezintă studiile anterioare. Practica sondajelor de opinie este una curentă, iar informații despre metodologiile utilizate în aceste anchete, instrumentul de colectare a datelor (chestionarul statistic), obiectivele și ipotezele de lucru, precum și despre rezultatele obținute există în abundență. Etapa de documentare presupune fișarea și structurarea acestor informații pentru a fi utilizate cu ușurință în elaborarea propriei anchete.

O altă sursă de documentare o reprezintă documentele oficiale din arhiva primăriei. Aceste documente sunt utile pentru a obține informații despre populația de referință, precum și date concrete despre activitatea primarului.

Interviurile exploratorii se pot realiza cu un număr de cetățeni ai orașului, cu angajați ai primăriei, cu analiști politici și ziariști.

d. Conceptualizarea

În această etapă, obiectivul principal este acela de a rezolva o serie de probleme punctuale privind abordarea teoretică cu care se va lucra asupra problemei de cercetat. Prin conceptualizare se poate înțelege etapa de analiză

conceptuală sau de analiză a conceptelor utilizate în cercetare și a teoriilor care privesc aceste concepte. Rezultatele obținute în etapa de explorare sunt utilizate aici pentru a evidenția:

- conceptele și ideile cheie ce privesc tema de cercetare;
- definițiile de lucru ale conceptelor utilizate, în special pentru conceptul de bază care definește tema de cercetare. Aceste definiții conferă un anumit sens conceptelor utilizate și pot fi strict teoretice sau operaționale;
- principalele teorii explicative utilizate în literatură pentru problema studiată. O teorie reprezintă un set de concepte, definiții, enunțuri care oferă o explicație asupra realității analizate, având un grad mare de generalitate și putere predictivă. Într-o abordare critică, se evidențiază plusurile și minusurile fiecărei teorii, care sunt cel mai frecvent utilizate în cercetările anterioare, cum a fost definit obiectul cercetării în aceste studii, care ar fi abordarea cea mai potrivită pentru actuala cercetare și de ce etc.;
- factorii determinanți care permit explicarea fenomenului, împreună cu ipoteze asupra tipului de impact al acestora;
- elementele teoretice conexe cu tema;
- ipotezele de lucru întâlnite în diverse studii;
- obiectivele de cercetare formulate în studii similare;
- dificultățile cercetării, problemele specifice, aspectele principale care definesc tema de cercetare etc.

Etapa finală a acestei etape de conceptualizare presupune o opțiune teoretică în care să se precizeze clar:

- conceptul de bază și alte concepte care susțin tema de cercetare;
- definițiile de lucru acceptate în cercetare pentru conceptele utilizate;
- orientarea teoretică aleasă: teoria cu care se lucrează în alegerea și definirea conceptelor cercetării;
- factorii determinanți și criteriile de clasificare a acestora;
- modelele explicative cu care se va lucra (modele teoretice, analitice etc.). Modelele sunt reprezentări schematice și simplificate ale realității care au rolul să prezinte caracteristicile principale ale fenomenului studiat și relațiile dintre diverse elemente.

Exemplu

Din punct de vedere teoretic, realizarea unui sondaj care analizează opinia populației privind activitatea unui primar și șansele de a fi ales din nou presupune o analiză conceptuală din domeniul teoriilor cu privire la opinia publică, în general, și cea politică, în particular. Literatura de specialitate pentru acest domeniu este foarte bogată, iar multitudinea cercetărilor prin sondaj realizate deja poate

reprezenta un avantaj, după cum poate constitui și o dificultate în formularea propriei problematici. Nu este cazul să prezentăm aici detalii cu privire la o problematică de cercetare pentru opinia populației. Câteva exemple care apar ca rezultate ale acestei etape și modul cum sunt utilizate în cercetare vor fi prezentate în legătură cu etapa de construire a unui model de analiză.

e. Construirea unui model de analiză sau a unui model de cercetare

Aceasta este etapa finală de definire și formalizare a temei de cercetare în care se concretizează întreaga muncă din etapele anterioare. Etapa presupune operaționalizarea conceptelor, definirea obiectivelor și formularea ipotezelor cercetării prin sondaj, validarea instrumentelor construite (Quivi și Van Campenhoudt, 1995; Mouton și Marais, 1990).

Operaționalizarea conceptelor

Problemei supuse cercetării prin sondaj i se asociază un *concept*, care reprezintă instrumentul teoretic de bază. Conceptul este o reprezentare cognitivă și simbolică a realității, un element produs prin abstractizare, ce fixează un anumit înțeles sau un conținut informațional și permite cercetătorului să creeze o relație cu realitatea supusă analizei. Pentru fiecare concept selectat se adoptă o definiție de lucru care precizează conținutul său informațional (definiție acceptată așa cum este prezentă în alte studii sau primind anumite nuanțe specifice abordării teoretice alese de cercetător). Prin munca de operaționalizare a conceptului de bază înțelegem efortul de a construi o serie de instrumente care să permită, în final, procesul de măsurare, adică observarea și exprimarea numerică a realității studiate (Mouton și Marais, 1990, p. 64). Conform literaturii de specialitate, operaționalizarea presupune identificarea principalelor *dimensiuni* ale conceptului de bază care instrumentează problema analizată; la nivelul fiecărei dimensiuni se identifică o serie de *componente*; pentru fiecare componentă se construiesc *caracteristicile* sau variabilele prin care se realizează efectiv măsurarea; în final, pe baza variabilelor, se definesc *itemii* și *întrebările* din chestionarul statistic. În etapa finală, pe baza chestionarului, se construiesc *variabilele* și *indicatorii* statistici care reprezintă purtătorii finali de informație, instrumente cu ajutorul cărora se ating obiective precise ale cercetării.

Dimensiunile conceptului. Un prim pas de trecere de la concept spre concretizarea procesului de măsurare îl reprezintă stabilirea dimensiunilor acestuia. Prin dimensiunile conceptului se identifică fenomene sau realități care apar ca elemente componente sau elemente de un grad mai redus de complexitate care se subsumează celei supuse analizei. Astfel, pentru conceptul de bază care definește tema de cercetare se identifică o structură conceptuală sau un set de concepte care se subsumează acestuia și care joacă rolul de a prelua o parte a problemei. Aceste

concepte cu un grad mai mic de generalitate poartă numele de *dimensiuni* și asigură împreună o structură conceptuală coerentă pentru fenomenul analizat. Unele concepte de bază permit ușor o asemenea structură, fiind obținute prin analiza modului de manifestare a fenomenului așa cum apare în realitate. De exemplu, pentru conceptele de educație, de venit se identifică ușor asemenea dimensiuni. Altele, însă, necesită o amplă dezbatere teoretică și admit mai multe posibilități de definire a dimensiunilor (de exemplu, conceptele de discriminare, violență, criză etc.).

Componentele. Deseori, trecerea de la dimensiuni la elementele operaționale ce permit observarea și măsurarea realității poate impune o altă structură intermediară, numită componentă. Practic, vorbim despre o structură conceptuală la nivelul unei dimensiuni, realizând același demers din etapa anterioară la acest nivel. Aceasta poate însă să lipsească și să se ajungă direct la variabile la nivelul unei anumite dimensiuni. Totul depinde de complexitatea temei și de posibilitățile de rafinare pe care le permite tema de cercetare. Componentele joacă același rol cu dimensiunile pentru conceptul de bază. Ideea este să se realizeze o detaliere suficientă a problemei analizate, atât în vederea unei mai bune înțelegeri, cât și a măsurării realității, a culegerii datelor.

Caracteristicile fenomenului se identifică la nivelul fiecărei componente sau dimensiuni. Aceste caracteristici reprezintă elemente observabile sau măsurabile ale unei anumite dimensiuni. Vorbim despre *variabile* sociale sau economice care permit observarea realității supuse cercetării. Operaționalizarea presupune un proces care conduce la ideea de a determina o listă sau un număr cât mai mare de caracteristici posibile, din care să fie alese ulterior cele mai semnificative în raport cu obiectivul principal al cercetării, precum și cu posibilitățile efective de culegere a datelor. În literatura de specialitate, aceste caracteristici sau atribute observabile sunt numite *indicatori*.

În următoarea etapă, se definesc elementele concrete care sunt elaborate în contextul construirii instrumentului de colectare a datelor. Mai precis, *itemii* și *întrebările* ce apar într-un chestionar statistic sunt construiți pe baza caracteristicilor stabilite pentru fiecare dimensiune.

În final, după stabilirea instrumentului de culegere a datelor, se definesc *variabilele* și *indicatorii statistici* utilizați în etapa de analiză a datelor. Variabilele statistice sunt instrumentele finale pe baza cărora se concretizează informațiile sau cunoștințele vizate în cercetarea științifică pe bază de sondaj. După tipul lor, acestea sunt numerice și nenumerice. Indicatorii statistici se referă la rezultate ale măsurării statistice și sunt purtători de informație, adică elemente de bază ale obiectivării statistice prin care se descrie și se explică realitatea studiată. Pentru

fiecare variabilă sau cuplu de variabile se delimitează indicatorii statistici care vor fi calculați pe baza datelor culese. Fiecare asemenea indicator este însoțit de metodologia necesară calculului și interpretării proprii.

În această lucrare preferăm o schemă a operaționalizării care este mai aproape de cercetarea cantitativă de tip statistic. Pentru un concept considerăm două posibile niveluri de descompunere ce rezultă din activitatea de analiză: dimensiunile și componentele. Urmează etapa operațională de identificare a caracteristicilor, ceea ce presupune definirea unor variabile reale concrete pentru fiecare dimensiune. În final, se construiesc instrumentele de culegere și analiză a datelor. Schematic, etapa de operaționalizare se poate prezenta în tabelul de mai jos. Ca exercițiu, pentru un concept de bază considerăm trei dimensiuni. Pentru prima și a treia dimensiune propunem câte două componente, iar pentru dimensiunea a doua, doar variabile.

Tabelul 1.1. *Schema operaționalizării conceptelor*

Concept				
Dimensiunea 1		**Dimensiunea 2**	**Dimensiunea 3**	
Componenta 11	**Componenta 12**	-	**Componenta 31**	**Componenta 32**
Variabila 111 Variabila 112	Variabila 121 Variabila 122 Variabila 123 Variabila 124	Variabila 21 Variabila 22	Variabila 311	Variabila 321 Variabila 322 Variabila 323
Item 1111 Item 1112 ...	Item 1211 Item 1221 ...	Item 211 Item 212 ...	Item 3111 ...	Item 3211 Item 3221 ...
Variabile statistice	Variabile statistice	Variabile statistice	Variabile statistice	Variabile statistice
Indicatori statistici	Indicatori statistici	Indicatori statistici	Indicatori statistici	Indicatori statistici

Exemplu

Operaționalizarea conceptului de opinie publică pentru ancheta propusă ca exemplu se poate realiza în mai multe moduri. Simplificând lucrurile, prezentăm mai jos o asemenea posibilă schemă de operaționalizare.

Concept: opinia cetățenilor orașului *A* cu privire la activitatea primarului și posibilitatea realegerii acestuia.

Dimensiuni: realizările din mandatul curent; caracteristici personale ale primarului; proiectul de viitor.

Componente

- pentru dimensiunea *realizări în mandatul curent* putem structura câteva componente care să identifice tipurile de realizări ale primarului pe domenii: administrativ, economic, social, politic etc.;
- pentru dimensiunea *caracteristici personale* se pot construi câteva componente: competențe manageriale, competențe de comunicare, calitatea morală etc.;
- pentru dimensiunea *proiectul de viitor* se pot identifica cel puțin două componente: opinia cetățenilor cu ce ar trebui realizat în oraș; opțiunea de vot pentru viitorul primar.

Caracteristici

De exemplu, pentru ultima dimensiune, componenta *opțiune de vot*, se pot construi variabile de genul: aprecierea primarului actual; aprecierea contra-candidaților; vot pentru primarul actual; vot pentru un alt candidat etc.

Itemi și întrebări

Un exemplu de item în chestionar pentru caracteristica *aprecierea primarului actual* este prezentat mai jos.

Apreciați în ce măsură sunteți de acord cu afirmația: rezultatele activității primarului localității în mandatul care tocmai se încheie sunt în conformitate cu promisiunile electorale.

1	2	3	4	5
Total dezacord				Total de acord

Variabile statistice

Pornind de la întrebarea prezentată mai sus, se poate defini variabila alternativă nominală care măsoară răspunsul persoanelor cu privire la prima opțiune de răspuns. Cele două valori posibile sunt *Da* și *Nu* și pot fi codificate cu valorile *1* și *0*.

Indicatori statistici

Continuând exemplul de mai sus, se poate stabili un indicator precum media variabilei alternative, adică proporția respondenților care au bifat prima opțiune de răspuns.

Definirea obiectivelor

Scopul principal al cercetării prin sondaj este producerea unei cunoaşteri ştiinţifice cu aplicaţii practice şi teoretice directe. Acesta este atins prin câteva categorii de obiective specifice cunoaşterii statistice:

- descrierea statistică a problemei la nivelul eşantionului extras;
- explorarea şi clarificarea anumitor dimensiuni ale problemei de cercetat;

- estimarea parametrilor distribuției populației totale după variabile de interes sau descrierea fenomenului la nivelul ansamblului;
- estimarea parametrilor modelelor care exprimă legătura dintre diferite variabile sau explicarea cauzelor fenomenului;
- prognoza unor evenimente viitoare;
- testarea ipotezelor statistice (se pot testa ipoteze: fie cu privire la parametri sau cu privire la legea de distribuție a populației totale după anumite variabile, caz în care se face apel la teoria testelor statistice parametrice; fie cu privire la structura populației sau la influența anumitor factori, caz în care se face apel la testele statistice neparametrice).

Toate aceste obiective sunt operaționale cu ajutorul variabilelor statistice și al indicatorilor aferenți prezentați în etapa de operaționalizare. Practic, fiecare indicator reprezintă instrumentul care permite obținerea unui rezultat sau a unei informații, adică permite specificarea unui obiectiv concret al cercetării.

Formularea obiectivelor cercetării în această variantă operațională și foarte precisă presupune și un exercițiu de evaluare și, eventual, de reformulare a unor obiective formulate în faza de debut a cercetării sau într-un document inițial elaborat de un beneficiar al cercetării.

Exemplu

Conform schemei de operaționalizare a conceptului de opinie, așa cum a fost prezentată anterior, pentru dimensiunea *proiect de viitor* s-au identificat componenta *opțiunea de vot*, caracteristica *vot pentru primarul actual* și indicatorul statistic *proporția cetățenilor care votează cu actualul primar*. În acest caz, se poate defini un obiectiv precis al cercetării, anume acela de a estima procentul persoanelor din oraș care ar vota din nou primarul actual la următoarele alegeri.

Elaborarea ipotezelor

În cadrul unei cercetări științifice, o ipoteză este o presupunere sau un enunț cu privire la tema studiată și care trebuie testată sau verificată cu ajutorul datelor culese conform metodologiei cercetării. De regulă, în enunțul unei ipoteze se specifică o relație între doi termeni bine delimitați, care apar fie ca realități concrete (fenomene, procese etc.), fie în concepte. În cele din urmă, prin etapa de operaționalizare, o ipoteză apare ca o relație între două sau mai multe caracteristici reale sau între variabilele statistice.

Ipotezele sunt formulate ca anticipări ale unor rezultate și de aceea iau forma unor enunțuri care conțin elemente precise, adică instrumentele identificate în etapa de operaționalizare. Astfel formulate, ipotezele pot fi testate în etapa de analiză statistică a datelor culese.

În orice cercetare prin sondaj se formulează un set de ipoteze care dau scheletul cercetării sau direcția etapei de analiză a datelor. Ca anticipări de rezultate, ipotezele se definesc în strânsă legătură cu obiectivele cercetării și cu structura conceptuală dezvoltată în etapa de operaționalizare. Astfel, ipotezele apar ca relații între dimensiunile și componentele conceptelor utilizate în anchetă, iar prin testarea acestora se obține acel tip de cunoaștere complexă specifică cercetării științifice.

Exemplu

O ipoteză ușor de construit pentru exemplul sondajului de opinie presupune o relație între realizările primarului actual și opțiunea de vot pentru viitorul primar: *Cetățenii satisfăcuți de ceea ce a realizat primarul orașului vor vota cu acesta la viitoarele alegeri.* O altă ipoteză ar putea fi formulată astfel: *Existența unor suspiciuni privind calitatea morală a primarului este strâns legată de opțiunea pentru un contracandidat.* În acest enunț se poate observa că se stabilește o relație între doi termeni precizați în etapa de operaționalizare: calitatea morală și opțiunea de vot pentru un nou mandat.

Concluzie

Modelul de analiză din cadrul unei cercetări științifice este reprezentat de un rezultat complex care așează împreună obiectivele cercetării rezultate în urma etapei de operaționalizare și setul de ipoteze de lucru ce stabilește relații între dimensiuni și componente ale conceptului studiat.

În tabelul 1.2 se prezintă un exemplu ipotetic de model de analiză ce apare sub forma unei scheme care identifică rezultatele etapei de operaționalizare și posibile relații dintre acestea, adică ipoteze ale cercetării. De asemenea, într-un asemenea tablou de sinteză se pot prezenta punctual și obiectivele cercetării identificate pentru fiecare dimensiune, respectiv pentru componentele adiacente.

Tabelul 1.2. *Model de analiză al cercetării*

Concept				
Dimensiunea 1		**Dimensiunea 2**	**Dimensiunea 3**	
		H_1		
		H_2	H_3	
Componenta 11	**Componenta 12**	-	**Componenta 31**	**Componenta 32**
		H_4		
Obiective	Obiective	Obiective	Obiective	Obiective

Validarea instrumentelor

Așa cum sugerează Mouton și Marais (1990), după etapa de operaționalizare se impune un exercițiu de evaluare a instrumentelor construite, pentru a răspunde la întrebarea dacă acestea sunt valide din punct de vedere științific. Cu alte cuvinte, acest efort presupune să evaluăm dacă instrumentele construite în această etapă corespund intenției de cercetare, adică într-adevăr măsoară ceea ce ne-am propus. Variabilele identificate în etapa de operaționalizare reprezintă instrumente de bază care sunt transpuse în instrumentele de culegere a datelor (întrebări și itemi în chestionarele statistice). Dacă aceste instrumente nu respectă anumite condiții de calitate, atunci etapele de culegere și analiză a datelor sunt viciate, iar rezultatele anchetei sunt afectate de erori.

Etapa de măsurare statistică presupune utilizarea unor instrumente de calitate. *Fiabilitatea* măsurării statistice (*reliability*) se referă la această proprietate de a asigura rezultate consistente prin repetarea procesului de observare. *Validitatea* procesului de măsurare se referă la gradul în care un instrument reușește să măsoare ceea ce a fost conceput să observe (cuantifice). Există mai multe metode de evaluare a calității instrumentelor și procesului de observare statistică. De exemplu, pentru un chestionar se utilizează analiza consistenței interne (coeficientul Conbach sau coeficientul „alphă", diverse teste de evaluare).

Validitatea reflectă acele erori de măsurare care apar sistematic în procesul de măsurare. În literatură se prezintă mai multe metode de evaluare a acestei proprietăți (Cozby, 2011; Mouton și Marais, 1990): *expert validity* (se realizează direct de cei interesați de rezultate); *content validity* (se verifică dacă elementele incluse în instrumentul de măsurare reprezintă în mod adecvat realitatea supusă observării); *criterion validity* (se utilizează un criteriu extern de validare a unei măsuri); *construct validity* (se verifică dacă o anumită măsură este consistentă cu conceptul teoretic supus evaluării).

1.3.2. Elaborarea chestionarului statistic

Problematica elaborării chestionarului statistic este una complexă și necesită mai mult spațiu decât un capitol într-o lucrare dedicată sondajului. Ca o sinteză, și, în același timp, ca un punct de plecare, pentru cei interesați să construiască un chestionar cu scopul de a fi aplicat într-o cercetare prin sondaj, prezentăm câteva elemente structurate pe trei niveluri de analiză: semantic (sens), sintactic (structură) și pragmatic (funcție). De asemenea, pentru abordarea propusă în acest capitol, utilizăm următoarea definiție de lucru: prin chestionar înțelegem o *succesiune logică și psihologică de întrebări, imagini și simboluri grafice* puse într-o formă

arhitecturală, care are funcţia de a determina un comportament asupra subiectului uman în legătură cu *obţinerea de informaţii* pe o anumită temă.

Definiţia prezentată mai sus sugerează cele trei dimensiuni fundamentale ale unui chestionar, în acord cu cele trei niveluri de analiză propuse (Jemna, 2004):

- conţinutul (întrebări, imagini şi simboluri grafice - nivelul semantic);
- forma sau maniera de aranjare a elementelor chestionarului într-un întreg coerent (nivelul sintactic). Nu orice succesiune de întrebări este un chestionar;
- funcţia (culegerea datelor în condiţii de calitate aşteptate - nivelul pragmatic).

În structura acestui subcapitol, potrivit metodei de analiză asumate, ne oprim la următoarele aspecte: analiza tipurilor de întrebări care compun un chestionar; arhitectura şi metodele de elaborare a chestionarului; tipuri de chestionare şi elemente specifice care pot înlesni realizarea funcţiei sale.

a. Construirea întrebărilor

Întrebările care apar în chestionar se pot clasifica după două criterii:

- în funcţie de posibilitatea de înregistrare a răspunsurilor, există întrebări: deschise, semi-deschise, închise;
- în funcţie de conţinutul lor, se identifică întrebări: de date factuale (ce face), de opinie (ce gândeşte), de intenţie (ce va face), de motivaţie (cum apreciază), care reclamă cunoştinţe (ce ştie), de identificare (cine este) etc.

În cazul *întrebărilor închise*, posibilităţile de răspuns sunt date sub forma unei liste complete, închise, persoana anchetată fiind constrânsă să aleagă unul sau mai multe din răspunsurile prezentate în chestionar.

Exemplu

Cum apreciaţi, în general, activitatea primarului localităţii *A* în ultimii 4 ani?

Foarte bună	
Bună	
Suficient de bună	
Satisfăcătoare	
Insuficientă	

În anchete se preferă întrebările închise pentru că sunt uşor de elaborat, de codificat şi, ulterior, facil de prelucrat. Problemele cele mai importante pe care le ridică întrebările închise sunt exhaustivitatea şi exclusivitatea listei răspunsurilor propuse (Lejeune, 1994, p. 66).

Într-o oarecare măsură, exhaustivitatea şi exclusivitatea sunt antinomice, căci, în tendinţa de a acoperi toate posibilităţile de răspuns, apar şi suprapuneri de sens. Este foarte greu de propus posibilităţi de răspunsuri mutual exclusive,

deoarece orice formulare, ca o construcţie semantică şi sintactică, este susceptibilă la interpretări în câmpul semantic al cuvintelor folosite. Dacă exhaustivitatea cere acoperirea tuturor posibilităţilor de răspunsuri, exclusivitatea cere o acoperire semantică între acestea. Din raţiuni de eficienţă, deseori se renunţă la exhaustivitate, după cerinţele anchetei, însă o listă exclusivă a răspunsurilor posibile este necesară pentru a înlătura atât redundanţele, cât şi neînţelegerile.

Exemplu

Care consideraţi că este cea mai importantă trăsătură personală a primarului localităţii *A*?

Modest	
Inteligent	
Bun orator	
Sociabil	
Atent	
Responsabil	
Parolist	

În exemplul de mai sus, lista răspunsurilor posibile este una foarte mare, astfel că exhaustivitatea nu poate fi atinsă. Lista poate fi mărită, în funcţie de interes, cu încă un număr de opţiuni, cu atenţie asupra condiţiei de exclusivitate.

În cazul *întrebărilor deschise*, există posibilitatea oricărui răspuns şi nu se impune nici o constrângere. Aceste întrebări sunt folosite în cazurile în care se doreşte să se obţină date sau informaţii mai sensibile sau în cazurile când elaborarea unei liste de răspunsuri nu este posibilă practic.

Exemplu

Care este motivul principal pentru care nu l-aţi mai vota pe actualul primar la viitoarele alegeri?

De regulă, întrebările deschise cer răspunsuri spontane, care redau cel mai bine latura subiectivă şi elementele de context. În acest sens, în constituirea întrebărilor deschise, un rol important îl ocupă stimulii. Prin stimuli se poate realiza mai uşor testarea unor comportamente şi obţinerea unor informaţii legate de atitudini, motivaţii, nevoi. Întrebarea deschisă este o structură cu caracter stimulativ, o structură „deschisă" cu implicaţie biunivocă: atât din partea expeditorului (cel ce concepe întrebarea), cât şi a destinatarului (cel căruia îi este adresată întrebarea). Astfel, formularea întrebării suportă exigenţa susceptibilităţii la interpretări, respondentul fiind plasat în faţa propriei limite de a formula şi a exprima un mesaj. În consecinţă, o anumită manieră de a folosi cuvintele corespunde unui anume mod de a gândi în societate. Iar aceste modele de comunicare, susţinute de ideologii, sunt mereu supuse transformărilor. Cazul României de după 1990 este semnificativ în acest sens. Foarte multe cuvinte au

suferit schimbări majore în câmpul lor semantic, iar folosirea lor cere cunoaşterea acestui sens de evoluţie (de exemplu, evoluţia cuvintelor din sfera politicului sau a economicului). Prin necesitatea cunoaşterii câmpului semantic al cuvintelor, întrebările deschise ridică probleme atât în faza de elaborare, cât mai ales în faza de analiză şi prelucrare. Aceste întrebări, care nu pot fi codate şi analizate după metodele cantitative, se pretează la o analiză calitativă, metoda cel mai des întâlnită fiind *analiza de conţinut*.

Întrebările semi-deschise au în lista răspunsurilor propuse şi o modalitate deschisă de răspuns.

Exemplu

Care dintre elementele specifice ale localităţii *A* a fost cel mai puţin folosit de actualul primar?

Resursa naturală	
Capitalul uman	
Poziţia geografică	
Altceva (ce anume?)	

Aceste întrebări rezolvă într-o anumită manieră problema exhaustivităţii întrebărilor închise pentru că oferă posibilitatea includerii unor răspunsuri diferite de cele din lista propusă. De asemenea, întrebările semi-deschise rezolvă în parte şi problema codării din cazul întrebărilor deschise. În faza de desprindere a datelor din chestionar, variantele de răspuns deschise pot fi codate, completând lista răspunsurilor propuse. O problemă a listei răspunsurilor o menţionează Claude Javeau, care arată riscul influenţei persoanelor prin sugestii la varianta deschisă (Javeau, 1978, p. 72). Este vorba despre un stimul arhitectural (lista răspunsurilor) care determină un comportament analog în cadrul modalităţii deschise. Prin varianta deschisă este posibil să nu se aducă nici un fel de contribuţie la cunoaştere.

Întrebările de date factuale sau de comportament sunt uşor de formulat şi de înţeles pentru cei intervievaţi. Răspunsul depinde de datele cerute, iar dacă se referă la comportamente sensibile, pot să apară erori prin efectul de dezirabilitate socială. Există o tendinţă a respondenţilor de a alege un răspuns care corespunde cel mai bine aşteptărilor celorlalţi sau unui statut social apreciat.

Exemplu

Ce aţi făcut la ultima acţiune organizată de primar privind curăţenia localităţii?

m-am alăturat celor care au cules deşeurile din parc	
am fost la plantat copaci	
mi-am văzut de treburile mele	
am fost doar să văd ce se întâmplă	

Întrebările de opinie pot ridica probleme atunci când persoana nu are o opinie clară sau nu are o opinie. În întrebările de opinie poate să apară problema falsei sinonimii a cuvintelor. Menționăm riscul de a se confunda opinia cu motivația. Pentru a stimula gândirea și formularea opiniei, se poate apela la formulări indirecte.

Exemplu

Care dintre domeniile de activitate considerați că au înregistrat o dezvoltare datorită implicării directe a primarului? Alegeți în ordine două variante.

Cultura	
Sănătatea	
Educația	
Serviciile	
Asistența socială	
Economia	

Întrebările de motivație. În comparație cu opinia (rezultatul unei analize critice, a unei evaluări axiologice), motivația reflectă implicațiile subiective ale persoanei în justificarea unei acțiuni, alegeri sau idei. În cercetarea motivației, problema cea mai delicată o constituie riscul sugerării unui răspuns. Dacă se produce acest risc, răspunsul persoanei nu oferă motivația reală, ci pe cea conștientizată în momentul anchetei.

Exemplu

La ultimele alegeri, în oferta electorală a primarului a existat și promisiunea egalizării șanselor între femei și bărbați. În ce măsură a cântărit acest lucru în decizia dvs. de vot?

1	2	3	4	5	6	7
În foarte mică măsură						În foarte mare măsură

Întrebările de intenție folosesc la studiul proiecțiilor, previziunilor personale. Răspunsurile la aceste întrebări pot fi distorsionate prin efectul dezirabilității sociale. O posibilitate de limitare a acestei erori ar fi formularea întrebării în variantă deschisă.

Exemplu

Ținând cont de spectrul politic actual, cu ce partid ați vota la alegerile viitoare?

PSD	
PNL	
PMP	
UDMR	
Altul (care?)	

Întrebările care reclamă cunoştinţe dau cea mai înaltă rată de non-răspunsuri. De obicei, sunt întrebări sensibile, însă pot juca un rol foarte important ca întrebări de control al consistenţei şi al consecvenţei răspunsurilor.

Exemplu

Care sunt cele trei obiective strategice pentru localitatea *A* anunţate de actualul primar în campania electorală?

Întrebările de identificare conţin, în general, date sensibile. Pentru a nu avea un efect de context, de obicei, sunt plasate la sfârşitul chestionarului. Aceste întrebări pot introduce erori sistematice (de exemplu, de rotunjire a vârstei). Pentru reducerea erorilor la întrebările care privesc vârsta, nivelul venitului sau alte date cifrice se recomandă utilizarea intervalelor de variaţie.

Exemplu

Starea civilă:

celibatar	
căsătorit	
divorţat	
văduv	

b. Arhitectura chestionarului

Subiectul arhitecturii unui chestionar aduce în discuţie cel puţin două probleme: tipurile de întrebări care joacă un rol sintactic în construcţia chestionarului; modalităţile de elaborare a unui chestionar.

Tipuri de întrebări în structura chestionarului

Septimiu Chelcea ne oferă o tipologie a întrebărilor din structura chestionarului care cuprinde: întrebări introductive, de trecere, filtru, de control şi de identificare (Chelcea, 1975, p. 187).

Întrebările introductive au rolul de interfaţă între persoana anchetată şi chestionar. Menţionăm că această funcţie o îndeplinesc cu mare succes semnele iconice. Prima pagină a chestionarului cuprinde, de obicei, sigla şi numele instituţiei care realizează ancheta. Aceasta captează atenţia şi oferă un prim impact cu destinatarul. Urmează o frază sau o imagine sugestivă care pune în temă interlocutorul cu privire la subiectul anchetei. Se pot aduce, în funcţie de context, informaţii suplimentare cu privire la motivul anchetei, la confidenţialitatea datelor etc. Toate acestea au rolul de a îndepărta sau canaliza diversele conotaţii care pot apărea la nivelul destinatarului cu privire la funcţia chestionarului. Se recomandă ca întrebările de introducere să fie închise; să nu fie întrebări sensibile sau la care este greu de răspuns. De exemplu, în introducere se pot folosi unele întrebări care au un interes general şi care vizează subiecte şi preocupări actuale. Prin aceste

întrebări se actualizează în mintea destinatarului măcar o parte a contextului temei cercetate. Există însă şi chestionare care, după o punere sumară în temă cu privire la anchetă, încep direct cu întrebări de identificare.

Întrebările de trecere delimitează structura chestionarului şi realizează legătura între diferite părţi ale acestuia, pregătind interlocutorul pentru ceea ce urmează. Putem vorbi aici despre coeziunea internă a chestionarului, care trebuie să fie puternică şi să dea o structură logică şi psihologică a conţinutului, pe linia realizării funcţiei chestionarului. Pentru obţinerea acestei coeziuni se folosesc elemente de trecere, de legătură, de relaţie. Acestea au rolul de a reduce proiecţiile efectelor stimulilor de la o secţiune la alta. Un exemplu de trecere, destul de evident în chestionar, este acela de la întrebările referitoare la tema anchetei la întrebările de identificare. Aceste întrebări pot fi prefaţate de câteva informaţii şi explicaţii care să reducă starea de susceptibilitate a destinatarului (de exemplu, o explicaţie a operatorului, în cazul chestionarului administrat, sau o frază, o imagine sau o figură, în cazul chestionarului autoadministrat).

Întrebările filtru supun unor condiţii necesitatea de a răspunde la o întrebare sau la un set de întrebări următoare. Aceste întrebări sunt deosebit de importante în cazul întrebărilor de date factuale. Prin întrebările filtru se evită răspunsuri care nu au acoperire în fapte.

Întrebările de control apar în structura chestionarului cu rolul de a verifica consecvenţa şi consistenţa răspunsurilor de-a lungul chestionarului. Întrebările de control apar în contexte diferite de subiectul lor. Prin întrebările de control se poate genera şi un comportament asupra persoanei anchetate, având un caracter stimulativ. Odată întâlnită o astfel de întrebare, persoana anchetată poate conştientiza faptul că nu trebuie să răspundă oricum.

Întrebările de identificare cer date şi informaţii cu privire la persoana celui anchetat. Ele se referă, cel mai adesea, la identificarea persoanei după sex, vârstă, nivel de pregătire, venit, stare civilă etc. Despre locul acestor întrebări în cadrul chestionarului, părerile sunt împărţite. Unii specialişti consideră că este bine ca ele să apară la sfârşit, pentru a nu avea un impact nefavorabil asupra persoanei chestionate, cunoscut fiind faptul că uneori se evită răspunsul la astfel de întrebări. Alţi specialişti consideră că întrebările de identificare trebuie doar să apară grupate, fie la începutul chestionarului, fie la sfârşitul lui, pentru a nu crea un efect de context şi a nu influenţa răspunsurile (Javeau, 1978, p. 78).

Tehnici de elaborare a chestionarului

În arhitectura chestionarului, pe linia mesajului, se poate vorbi despre o curbă de tensiune care exprimă intensitatea relaţiei dintre emiţător şi receptor de-a lungul chestionarului. Pentru a face posibilă funcţia chestionarului, structura sa

trebuie să aibă caracter stimulativ în această direcţie. În consecinţă, începutul chestionarului trebuie să capteze atenţia. Primele elemente din chestionar sunt construite pentru a suscita un interes mare pentru destinatar, chiar dacă la nivelul expeditorului mesajul are interes scăzut. Pe măsură ce se parcurge chestionarul, raportul se inversează. Interesul destinatarului începe să scadă, iar cel al expeditorului atinge valoarea maximă spre finalul chestionarului. Între cele două extreme, se poate vorbi despre o relaţie strânsă, generată de un interes comun şi un raport de tensiune proporţional. Pentru a realiza această curbă de tensiune, în funcţie de tema anchetei şi de populaţia de referinţă, se pot utiliza următoarele tehnici de construire a chestionarului (Chelcea, 2004, p. 131): tehnica pâlniei, tehnica pâlniei răsturnate, tehnica ce ţine cont de efectul „halo" .

Tehnica pâlniei

Are la bază principiul deducţiei, adică trecerea de la general la particular. Chestionarul se deschide cu întrebări de ordin general, care sunt, de obicei, deschise şi îi lasă interlocutorului posibilitatea exprimării subiective cu privire la tema propusă. Această arhitectură are caracter stimulativ şi are scopul de a reduce influenţa anumitor tipare, coduri, mentalităţi etc. Cu fiecare întrebare se construieşte un cadru care canalizează atenţia spre problemele specifice, particulare ale temei, probleme de mare importanţă pentru anchetă. Tehnica pâlniei dă rezultate în cadrul populaţiilor cu un nivel de pregătire mai ridicat (la persoane care pot să răspundă la întrebările de ordin mai general, fără a avea un comportament de respingere) şi în cazul unor teme delicate, care necesită o pregătire a persoanei anchetate. Pentru populaţiile cu un nivel de pregătire mai scăzut se recomandă tehnica bazată pe principiul opus.

Tehnica pâlniei răsturnate

Are la bază principiul inducţiei, adică drumul de la particular la general. Chestionarul debutează cu întrebări închise, de date factuale sau de opinie, la care se poate răspunde uşor. Prin această tehnică, stimulul arhitectural are rolul de a facilita persoanei procesul de înţelegere şi de comunicare la nivel sintetic, general.

Tehnica ce ţine cont de efectul „halo"

Prin această tehnică se are în vedere comportamentul sentimental sau logic al persoanei la impactul cu anumite întrebări sensibile ale chestionarului. Impactul poate determina o influenţare sau o contaminare a răspunsurilor ulterioare (efect „halo"). Corelaţiile pe care le face persoana anchetată între subiectul întrebării (sau tema anchetei) şi fondul său cultural sau sentimental generează conotaţii emoţionale sau ideologice. De exemplu, în cazul anchetelor cu caracter politic, religios, sentimental sau cu alte asemenea teme sensibile, efectul „halo" poate avea efecte semnificative. Evitarea acestor efecte se poate realiza prin aşezarea

întrebărilor sensibile fie la finalul chestionarului, fie într-un cadru în care sunt pregătite prin întrebări de trecere, semne iconice şi informaţii explicative.

c. Calitatea chestionarului. Probleme specifice

Funcţia chestionarului presupune culegerea de date de la populaţia anchetată în condiţii de calitate ce presupun erori cât mai mici, timp de completare cât mai scurt, costuri reduse. În funcţie de tema de cercetare, de populaţia de referinţă şi de metoda de sondaj aleasă, se poate defini o tipologie a chestionarelor statistice. Pentru fiecare caz în parte, atingerea funcţiei chestionarului presupune o serie de condiţii de calitate particulare. Pe lângă tipul de chestionar ales, este important să se analizeze o serie de probleme specifice procesului de elaborare, astfel încât produsul finit să faciliteze cel mai bine culegerea datelor. Ne referim aici la lungimea chestionarului, editare, testare etc.

Tipuri de chestionare

Tipologia chestionarului ne oferă un instrument metodologic cu ajutorul căruia putem evidenţia domeniile de utilizare a chestionarului, elementele componente, posibilităţile sale de aplicare etc. O abordare clasică a tipologiei chestionarului ne este oferită de Septimiu Chelcea, care propune următoarele criterii de clasificare: conţinutul informaţiilor, forma întrebărilor, modul de aplicare a chestionarului (Chelcea, 1975, p. 143).

După conţinutul informaţiilor, există două tipuri de chestionare: chestionare de date factuale şi chestionare de opinie.

Chestionarele de date factuale sunt chestionare de tip oficial şi au un scop administrativ. Aceste chestionare sunt de cele mai multe ori tipizate şi folosesc la culegerea unor date periodice. Datele culese cu aceste chestionare au un caracter obiectiv şi se referă la fapte.

Chestionarele de opinie se adresează opiniilor şi atitudinilor personale, iar informaţiile culese cu ajutorul lor au îndeosebi un caracter subiectiv. Pe lângă opinii, motivaţii, atitudini, comportamente, prin acest tip de chestionare se obţin şi date de identificare a persoanei, date care au un caracter obiectiv. În general, chestionarele de opinie actuale sunt construite cu întrebări care privesc ambele tipuri de date, atât fapte, cât şi atitudini, opinii, iar întrebările de identificare sunt prezente în chestionar pentru a corela datele şi informaţiile obţinute cu profilul personalităţii celui care răspunde.

După forma întrebărilor, pot fi: chestionare cu întrebări închise, chestionare cu întrebări deschise şi chestionare cu întrebări mixte.

Chestionarele cu întrebări închise sunt chestionarele în care răspunsurile la întrebări sunt fixate dinainte, iar persoana care va răspunde trebuie să aleagă un răspuns din cele propuse. Întrebările închise dau posibilitatea codării în momentul

elaborării, iar chestionarul cu astfel de întrebări se mai numeşte şi chestionar precodat. Precodarea presupune atribuirea unui cod numeric sau alfabetic fiecărui răspuns propus şi dă posibilitatea unei prelucrări ulterioare mult mai facile.

Chestionarele cu întrebări deschise conţin întrebări care nu oferă variante de răspuns; răspunsurile sunt lăsate la latitudinea celui chestionat.

Chestionarele cu întrebări mixte au în componenţa lor atât întrebări închise, cât şi întrebări deschise. Acestea sunt chestionarele cel mai des folosite.

După modul de aplicare a chestionarului, se practică două tipuri: chestionare autoadministrate şi chestionare administrate de un anumit personal (operatori).

Chestionarele autoadministrate sunt numite aşa pentru că răspunsurile la întrebări sunt date direct de persoana interogată, fără nici o mediere din partea anchetatorului. Cele mai folosite tipuri de chestionare autoadministrate sunt cele trimise prin poştă, cele publicate în ziare, reviste şi pe internet, cele care însoţesc diferite produse sau care se găsesc în magazine.

Chestionarele administrate sunt completate de un personal specializat, prin dialog direct cu persoanele anchetate.

Lungimea chestionarului

Problemele legate de lungimea chestionarului se referă la câteva aspecte. În primul rând, se are în vedere nevoia de informaţie şi capacitatea de a sintetiza această nevoie într-un număr restrâns de întrebări. În al doilea rând, există o constrângere de ordin economic, care se referă la costuri. Urmează aspectele legate de populaţia anchetată şi de condiţiile de administrare, care impun o limită de timp. De exemplu, ca durată de timp, o anchetă realizată pe stradă nu poate fi mai lungă de zece minute, iar una efectuată la domiciliu nu poate depăşi 30-40 de minute. Dacă chestionarul este lung şi complex, pot apărea erori datorate: fie oboselii, caz în care scade capacitatea de concentrare şi pot apărea răspunsuri eronate, fie lipsei de timp sau de disponibilitate, care duce uşor la non-răspunsuri. În unele cazuri, chiar dacă chestionarul este mai lung, dar nu peste o anumită limită, aceste reacţii comportamentale pot fi ameliorate. Orice relaţie cere un anumit interes. Dacă uneori este posibil să se stimuleze acest interes pe cale materială, mult mai important este ca chestionarul însuşi să suscite interes. O astfel de situaţie presupune ca forma şi conţinutul chestionarului să fie interactive, stimulative, tensionale. Elementele chestionarului trebuie să „curgă" într-o anumită coerenţă logică şi psihologică ce poate să întreţină relaţia până la capăt, fără ca lungimea chestionarului să fie un impediment.

Editarea chestionarului

Dacă forma şi structura chestionarului ne dau arhitectura sa conceptuală care face posibilă funcţia, editarea sau punerea în pagină constituie materializarea

acestei structuri şi reprezintă un element de interfaţă. Prin editare se adună toată munca „nevăzută" de elaborare a unui chestionar. Caracterul său este stimulativ, atât în cazul chestionarului administrat de operator, cât şi în cazul chestionarului autoadministrat. În ipoteza unui conţinut şi a unei structuri conceptuale de calitate, o materializare care nu este pe măsură poate diminua eficienţa chestionarului. Designul chestionarului cere, în primul rând, vizibilitatea formei. Claritatea, densitatea şi aspectele de simetrie în delimitarea elementelor chestionarului, apoi culoarea şi estetica formelor, calitatea imprimării şi a hârtiei, toate sunt elemente importante care contribuie la reducerea erorilor de observare. Importanţa prezentării materiale a chestionarului devine mult mai evidentă în cazul chestionarului autoadministrat. Un chestionar de o bună calitate editorială impune seriozitate, profesionalism şi contribuie la câştigarea încrederii persoanei anchetate.

Testarea chestionarului

Înainte de aplicarea chestionarului, are loc testarea acestuia pe un eşantion mic, de maxim 50 de persoane, în condiţii cât mai apropiate de ancheta propriu-zisă. Interesul testării este să evidenţieze dacă chestionarul oferă informaţia dorită, precum şi dificultăţile care apar în momentul culegerii datelor. Ameliorarea chestionarului se poate realiza prin modificările sugerate de testul efectuat. În faza de testare, administrarea chestionarului se poate transforma, ulterior, într-un scurt interviu. Se pot cere explicaţii despre răspunsurile oferite şi despre modul de înţelegere a întrebărilor, despre logica şi coerenţa întrebărilor în chestionar etc. În faza de testare a chestionarului, sunt importante răspunsurile la următoarele întrebări:

- sunt înţelese corect întrebările?
- se dau mai multe răspunsuri diferite la aceeaşi întrebare în condiţii similare?
- care întrebări provoacă stimuli şi reacţii diferite?
- informaţiile cerute sunt la dispoziţia persoanelor anchetate sau este nevoie de un efort dificil de amintire?
- vocabularul este adecvat persoanelor anchetate şi problemei studiate?
- listele cu răspunsuri sunt exhaustive şi exclusive?
- ordinea întrebărilor este logică?
- există efect de context?
- care întrebări sunt sensibile şi ce efect au?
- etc.

Informaţiile obţinute prin testare pot contribui la ameliorarea calităţii chestionarului prin câteva modificări de conţinut sau de formă: se pot modifica anumite întrebări, de exemplu, unele întrebări deschise pot fi transformate în

întrebări închise, se pot înlocui cuvintele ambigue sau abstracte, se pot reformula întrebările sensibile, se poate schimba ordinea anumitor întrebări etc.

1.3.3. Elaborarea planului de sondaj

Planul de sondaj vizează un ansamblu de instrumente și procedee care definesc principiul și modalitatea de extragere a unei unități din populație pentru a intra în eșantion, structura și condițiile de calitate ale eșantionului. Elaborarea unui plan de sondaj presupune:

- definirea populației de respondenți - stabilirea caracteristicilor de omogenitate a populației;
- identificarea informațiilor suplimentare și a surselor acestora;
- definirea bazei de sondaj - o listă cu unitățile individuale ale populației (când acest lucru este posibil);
- alegerea unei metode de sondaj - aleatoare, empirice, mixte;
- definirea eșantionului - volum, structură, metodă de extragere;
- stabilirea gradului de precizie a rezultatelor.

a. Definirea populației de referință. Informații suplimentare

Așa cum am stabilit în secțiunea anterioară, în sondajele statistice ne interesează să stabilim o populație statistică ce vizează respondenții sau cei care oferă informațiile în procesul de culegere a datelor. Așadar, în prima etapă a planului de sondaj se definește această populație sau colectivitate de referință. În general, o populație statistică este definită prin precizarea a patru factori: natura, caracteristicile intrinseci, spațiul și timpul (Ardilly, 1994, p. 4).

Natura populației determină natura și modul de organizare a unităților sale, care pot fi: simple (persoane) sau complexe (firme, familii, unități administrative etc.). În această etapă este important să se definească și elementele de identificare ale unităților individuale din populație. Distincția dintre o unitate individuală de eșantionare și cea de observare, realizată în secțiunile anterioare, este, de asemenea, foarte importantă. Această distincție se referă la identificarea unităților individuale potrivit naturii populației (unitățile de eșantionare se definesc după acest criteriu al modului de organizare a populației) și a posibilităților de culegere a datelor (unitățile de observare sunt cele care participă efectiv la culegerea datelor). În cele mai multe cazuri, cele două coincid, însă există și situații când diferă (de exemplu, în anchetele legate de bugetele de familie, unitatea individuală este familia, dar cea de observare este o anumită persoană din familie, care trebuie bine desemnată).

Caracteristicile intrinseci (de exemplu, înălțimea, cifra de afaceri etc.) joacă un rol important în stabilirea omogenității populației și a posibilităților de extragere

a eşantionului. Spaţiul sau localizarea (de exemplu, municipiul Iaşi, judeţul Iaşi, regiunea de Nord-Est a României etc.) asigură delimitarea geografică a fenomenului. Data sau timpul (de exemplu, la 01.04.2018, în perioada 01.03 - 01.04.2018) este un factor decisiv în definirea populaţiei. Dacă definirea populaţiei se face pe baza unor date şi informaţii dintr-o cercetare anterioară (de obicei, dintr-un recensământ), există posibilitatea apariţiei erorilor de neconcordanţă între populaţia reală şi cea definită prin planul de sondaj.

b. Informaţii despre populaţie

Informaţiile suplimentare care se pot obţine privesc în special structura populaţiei şi gradul de omogenitate a acesteia. Clasificarea populaţiei pe grupe cu omogenitate mai ridicată decât întregul reprezintă o soluţie importantă pentru creşterea gradului de precizie a rezultatelor. Această grupare presupune cunoaşterea structurii populaţiei pentru una sau mai multe variabile puternic corelate cu tema de cercetare. Existenţa acestei informaţii este determinantă pentru alegerea unei metode de sondaj. Utilă este şi informaţia privind gradul de omogenitate a populaţiei în raport cu o variabilă de interes sau cu o variabilă corelată cu tema de cercetare. Această informaţie este esenţială pentru calculul volumului eşantionului în cazul sondajelor aleatoare sau mixte.

c. Construirea bazei de sondaj

Baza de sondaj reprezintă o listă a unităţilor individuale ale populaţiei sau o bază de date care cuprinde elementele de identificare a unităţilor populaţiei după anumite caracteristici. O bază de sondaj trebuie să aibă câteva caracteristici principale: să permită identificarea clară, fără nici o problemă a unităţii individuale din populaţie, să fie completă sau exhaustivă, să nu fie redundantă, să ofere posibilitatea unei gestiuni facile (Ardilly, 1994, p. 24).

Identificarea unităţilor individuale se face pe baza unor caracteristici numite de identificare. De exemplu, pentru o persoană, identificarea se face pe baza numelui şi a adresei domiciliului stabil, a unui cod (CNP sau cod de securitate socială etc.). Condiţia de exhaustivitate este importantă pentru selecţie. Dacă din baza de sondaj lipsesc unităţi care, în mod real, fac parte din populaţia de referinţă, se comite o eroare numită eroare de acoperire. În acest caz, se modifică şi probabilităţile de includere a unităţilor în eşantion. Redundanţa presupune ca baza de sondaj să nu aibă duble înregistrări, fapt care ar conduce la eroarea de a nu asigura aceeaşi şansă fiecărei unităţi de a intra în eşantion. Facilitatea gestiunii priveşte atât extragerea eşantionului, cât şi utilizarea unor informaţii suplimentare utile în faza de ameliorare a rezultatelor obţinute în sondaj.

Baza de sondaj poate fi formată din unităţi individuale, cum ar fi persoanele, sau din unităţi complexe, cum ar fi: ariile geografice, unităţile administrativ-

teritoriale, clădirile, firmele etc. Avantajul bazelor de date compuse din unităţi complexe este, în primul rând, unul de stabilitate şi poate asigura construcţia unor planuri de sondaj care iau în considerare modul de organizare a populaţiei (de exemplu, sondajele pe serii sau sondajele în mai multe trepte).

Pentru multe situaţii practice care necesită organizarea unui sondaj, este dificil sau chiar imposibil de construit o bază de sondaj. În asemenea cazuri, se realizează sondaje empirice, mixte, sondaje în mai multe trepte, sondaje de tip cluster. De exemplu, există baze de sondaj, cum ar fi cele de la evidenţa populaţiei, însă nu sunt accesibile şi nu pot fi utilizate ca sursă de informare. De asemenea, există baze de date ale unor instituţii precum Finanţele Publice, Camera de Comerţ şi Industrie, administraţiile publice locale, firmele de distribuţie a utilităţilor. În general, aceste baze sunt construite regional şi necesită o integrare (dar, şi în aceste cazuri, datele nu sunt accesibile pentru organizarea de sondaje). Mai pot exista baze de date construite de diverse instituţii, cum ar fi Institutul Naţional de Statistică, instituţiile private de sondaje sau instituţiile administrative (cele care construiesc listele electorale), însă problemele acestor baze sunt actualitatea lor (incluzând şi posibilitatea de actualizare a datelor) şi costul ridicat al informaţiei.

d. Alegerea metodei de sondaj

În lucrările de teoria şi practica sondajelor, în funcţie de principiul de selecţie a unei unităţi în eşantion, se delimitează două metode de sondaj: *probabiliste* sau aleatoare şi raţionate sau *empirice*. O combinaţie a celor două metode de bază conduce la o a treia categorie - *sondajele mixte* - care sunt cele mai utilizate în practică.

Sondaje aleatoare

Sondajele aleatoare permit calcularea a priori a probabilităţii fiecărei unităţi din populaţie de a aparţine eşantionului. Acest lucru presupune existenţa unei baze de sondaj şi a unei scheme probabiliste de extragere. Prin această metodă se asigură baza de calcul a erorilor de eşantionare, a volumului eşantionului şi determinarea preciziei rezultatelor. În funcţie de şansa fiecărei unităţi de a intra în eşantion, se poate aplica fie sondajul aleator cu probabilităţi egale (sondaj aleator simplu), fie sondajul aleator cu probabilităţi inegale. Există situaţii când aceste sondaje se pot aplica practic ca atare, potrivit metodologiei specifice sondajului aleator, dar, frecvent, acest principiu de selecţie este utilizat într-o anumită etapă a unui sondaj mai complex.

Sondajele empirice

Metodele de sondaj empirice se bazează pe o serie de criterii de selecţie care depind de natura populaţiei şi de informaţiile disponibile. Principiul care justifică selecţia unităţilor are o bază empirică şi vizează obţinerea unui eşantion pe baza

unei relații de corespondență cu întregul, potrivit unui criteriu bine stabilit (*choix raisonné*). Aceste metode se folosesc atunci când este dificil (economic și științific) de realizat un sondaj aleatoriu sau mixt.

Obiectivul acestei metode este obținerea de rezultate cât mai aproape de un sondaj aleator. Reprezentativitatea este asigurată printr-o corespondență între populația totală și eșantion, fie la nivelul structurii, fie după alte criterii care asigură procesul de selecție.

Cel mai des întâlnite metode de sondaj empirice sunt:

- metoda sondajului pe cote, care presupune obținerea unui eșantion ca o miniatură a populației după o structură dată de anumite variabile fundamentale pentru anchetă, numite variabile de control;
- metoda unităților tip, care are ca principiu de bază împărțirea populației în categorii omogene și selectarea în eșantion a unităților care sunt cel mai aproape de media fiecărei categorii sau sub-populații;
- metoda itinerariilor, o variantă a metodei cotelor, prin care operatorilor li se impune mult mai strict modalitatea de alegere a unităților care participă la anchetă, pe baza unui itinerar (un ansamblu de adrese, străzi sau zone din care se extrag persoane după anumite consemne clare).

Sondajele mixte

Combinarea metodei aleatoare cu cea empirică presupune obținerea unei metode de sondaj mai eficiente. Criteriul de eficiență vizează atât precizia rezultatelor, cât și dimensiunea economică și de timp. Metodele mixte presupun combinarea unor principii sau instrumente oferite de cele două metode de bază în diferite forme și în diferite etape ale procesului de eșantionare. Pentru exemplificare, facem trimitere la trei asemenea metode mixte: sondajul stratificat, în mai multe trepte, pe grupe (cluster).

Sondajul *stratificat* admite, mai întâi, o componentă empirică ce presupune stratificarea populației pe grupe omogene cu ajutorul unei variabile corelate semnificativ cu tema analizată. În pasul al doilea, din fiecare strat se extrag aleator sub-eșantioane de unități care vor constitui, în final, eșantionul anchetei.

În cazul sondajului în *mai multe trepte* se aplică principiul stratificării de mai multe ori, la mai multe niveluri. De exemplu, într-o primă treaptă, se constituie o structură a populației dintr-un număr de unități complexe primare asemenea straturilor. La acest nivel, are loc o extragere aleatoare a unui număr de unități primare. Ulterior, din fiecare unitate primară selectată urmează o altă selecție aleatoare a câte unui sub-eșantion de unități complexe secundare sau individuale. Procesul se poate realiza în două sau mai multe etape.

Sondajele pe *serii* (*cluster*) presupun utilizarea unor structuri naturale ale unităților populației numite serii sau clustere. În acest caz, selecția aleatoare presupune extragerea unui număr de serii din populație, iar eșantionul final se constituie din toate unitățile seriilor extrase.

Criterii de alegere a unei metode de sondaj

Opțiunea pentru una dintre cele două metode de sondaj, precum și îmbunătățirea eficienței fiecărei metode, se realizează în funcție de mai multe criterii[2]: informația disponibilă, precizia, costul, timpul, natura fenomenului, volumul eșantionului.

Informația disponibilă

Pentru populația de referință se pot obține informații de diverse tipuri, prin intermediul mai multor surse: studii anterioare, documente, cercetări preliminare anchetei etc. Aceste informații sunt disponibile fie la nivel de unitate individuală, fie la nivelul întregii populații și pot fi utilizate în faza de obținere a eșantionului (a priori) sau după realizarea sondajului (a posteriori). Dacă la nivelul populației totale se cunoaște o informație dată printr-o variabilă puternic corelată cu variabila de interes, se poate realiza un plan de sondaj care respectă condițiile de eficiență: este mai precis, se realizează mai repede, necesită un eșantion de volum mai redus. Un exemplu este sondajul stratificat sau sondajul pe cote. De asemenea, o informație suplimentară se poate utiliza după ce a fost extras eșantionul, conform unui plan de sondaj, pentru a ameliora calitatea rezultatelor. Decizia de a utiliza informația suplimentară în construirea unui plan de sondaj sau în faza de ameliorare a rezultatelor se ia în funcție de celelalte criterii utilizate în analiza calității unui sondaj.

Precizia rezultatelor

Pentru calitatea unui sondaj, precizia reprezintă criteriul fundamental. În practică, acest criteriu este frecvent pus în balanță cu cel economic sau de timp, iar alteori se ia decizia de a realiza un tip de sondaj impus de natura fenomenului, fără a putea stabili o condiție de precizie. Totuși, cel mai frecvent, se poate impune o limită a preciziei în funcție de alte criterii, cum ar fi costul și timpul de realizare ale sondajului. Pentru sondajele aleatoare și mixte se poate determina gradul de precizie a rezultatelor, în timp ce pentru sondajele empirice se pot face doar aprecieri bazate pe expertiză. Sondajele mixte sunt mai precise decât cele aleatoare simple.

---

[2] Am prezentat asemenea criterii în lucrarea: D.V. Jemna, „Criterii de alegere a unei metode de sondaj în cercetarea statistică", *Analele Universității „Al.I. Cuza" din Iași*, Tomul LII/LIII, 2005-2006, pp. 373-379.

Costul

Marea majoritate a sondajelor sunt realizate în condiții de cost fixate a priori. Cu alte cuvinte, costul total al anchetei poate reprezenta o restricție importantă pentru a decide în privința unui tip de sondaj. Realizarea unui sondaj nu este doar o activitate științifică, ci și o activitate economică. Dintre mai multe planuri de sondaj cu un grad de precizie apropiată, de obicei, se optează pentru un plan de sondaj mai economic, dacă eforturile sunt semnificativ mai reduse. Sondajele empirice sunt, de obicei, mai puțin costisitoare, în comparație cu cele aleatoare sau mixte. Sondajele aleatoare presupun, de regulă, un eșantion de volum mare și existența unei baze de sondaj, condiții ce conduc la creșterea costului anchetei. În condițiile în care costul anchetei nu reprezintă o problemă, dacă situația o permite, este preferabil să se organizeze un sondaj aleator sau mixt.

Timpul de realizare

În cele mai multe situații practice, rezultatele unui sondaj trebuie obținute într-un timp cât mai scurt, pentru că aceste rezultate fundamentează decizii concrete cu privire la fenomenul studiat. Timpul de realizare se diminuează în contextul anchetelor care se realizează cu o anumită frecvență (utilizând metodologia, infrastructura și rezultatele din anchetele anterioare). Sondajele empirice se pot realiza într-un timp mult mai scurt, în comparație cu cele aleatoare sau mixte. Criteriul timp poate fi decisiv, uneori, pentru alegerea unei metode de sondaj (caz în care se poate apela la o metodă empirică).

Natura populației de referință

Deși gradul de precizie, costul și timpul de realizare a sondajului sunt criteriile care definesc eficiența unei anchete, acestea nu pot fi evaluate fără a ține seama de natura populației. Acest criteriu poate restrânge plaja opțiunilor pentru un anumit tip de sondaj. Există situații când natura fenomenului sau a populației de referință impune cu necesitate un anumit tip de sondaj (de exemplu, în situația în care nu există o bază de sondaj sau nu se cunosc informații suplimentare pentru populația de referință, când volumul populației este necunoscut sau unitățile individuale se delimitează doar în momentul observării).

Volumul eșantionului

Cel puțin la nivel intuitiv, se poate considera că un sondaj este cu atât mai bun cu cât eșantionul este mai mare, dacă se respectă condițiile de reprezentativitate. Mărimea eșantionului este legată de metoda de sondaj aleasă și determină precizia rezultatelor, cel puțin până la un anumit nivel. În cazul sondajelor aleatoare, volumul eșantionului reprezintă o condiție importantă pentru precizia rezultatelor. De regulă, pentru aceste sondaje se impune o limită de precizie care determină volumul eșantionului. În general, sondajele aleatoare

presupun eșantioane de volum mare, ceea ce implică costuri ridicate și un timp de realizare mai lung. Sondajele mixte, prin utilizarea unei informații suplimentare, pot reduce semnificativ volumul eșantionului, însă limita este stabilită de nivelul de precizie sau de cost. În cazul sondajelor empirice, nu există un criteriu științific de definire a volumului eșantionului. Limitarea acestuia reprezintă o condiție impusă de cost și de timpul de realizare.

e. Definirea eșantionului

Eșantionul reprezintă acea sub-populație extrasă din populația de referință conform metodei de sondaj alese și care respectă principiul reprezentativității. Pentru un eșantion statistic sunt importante două caracteristici (volumul și structura acestuia) care pot fi definite în contextul fiecărui tip de metodă de sondaj.

Cazul sondajului aleator

Eșantionul care rezultă în acest caz respectă în principal o condiție de volum. Volumul eșantionului trebuie cunoscut a priori. Acesta se calculează în condiții de probabilitate asumate și a unei erori de estimare fixate. Pentru eșantionul care va fi extras nu se impune o anumită structură, ci doar respectarea principiului selecției aleatoare, adică a extragerii unei unități cu probabilitatea de includere cunoscută și după o schemă de extragere probabilistă. Pentru sondajele aleatoare simple, probabilitatea de includere este aceeași pentru toate unitățile din populație. Extragerea eșantionului se realizează cu ajutorul unui algoritm și al unui generator de numere aleatoare.

Cazul sondajului empiric

Pentru aceste tipuri de sondaje, volumul eșantionului nu se poate calcula sau estima. În funcție de tipul de sondaj empiric aplicat, specialiștii propun utilizarea unui volum al eșantionului care este stabilit fără apelul la metode și principii de calcul, ci la expertiză și practici anterioare. Eșantioanele obținute în acest caz au, de regulă, o structură bine stabilită în faza de elaborare a planului de sondaj. De exemplu, sondajul pe cote utilizează o serie de variabile care structurează populația, iar selecția presupune obținerea unui eșantion cu o structură identică. Pentru că volumul eșantionului nu se poate calcula, în dimensionarea acestuia se apelează la opinia experților, la studiile realizate anterior și la informațiile disponibile despre populația de referință.

Cazul sondajului mixt

Sondajul mixt permite atât calculul volumului eșantionului, cât și determinarea structurii acestuia. De exemplu, în cazul sondajului stratificat, eșantionul păstrează structura populației de la nivelul straturilor, iar din fiecare strat se extrag aleator un număr de unități care se calculează a priori.

f. Erorile de eşantionare şi precizia rezultatelor

Erorile de eşantionare şi precizia rezultatelor pot fi tratate separat în funcţie de principiul de selecţie utilizat: aleator şi empiric.

Sondajul aleator şi mixt oferă posibilitatea măsurării erorilor de eşantionare. Instrumentele care asigură măsurarea acestor erori, precum şi extinderea rezultatelor de la eşantion la populaţia totală, sunt *estimatorii*, adică variabile aleatoare construite pe mulţimea tuturor eşantioanelor posibil de extras dintr-o populaţie. Proprietăţile estimatorilor asigură posibilitatea măsurării erorilor de estimare, precum şi realizarea unor teste care privesc calitatea rezultatelor sondajului.

Precizia rezultatelor obţinute în sondajele empirice nu poate fi apreciată decât intuitiv, pe baza criteriilor care au stat la baza eşantionării. Se poate realiza o precizie ridicată dacă populaţia este suficient de bine cunoscută după variabilele de control şi dacă eşantionarea se realizează în condiţii de calitate. În acest caz, decisivă este echipa care realizează eşantionarea şi culegerea datelor. O cât mai bună instruire a operatorilor de teren şi o cât mai detaliată procedură de identificare a unităţilor care compun eşantionul asigură şi şansele de a obţine rezultate cu erori cât mai mici posibile.

1.4. Probleme organizatorice

Realizarea unei cercetări pe bază de sondaj statistic nu se limitează la construirea unui model de analiză şi a unui chestionar de culegere a datelor, ci presupune şi un efort organizatoric, care este unul foarte important pentru succesul cercetării. Dintre activităţile ce se referă la etapa de organizare a unui sondaj, selectăm următoarele:

- stabilirea necesarului uman, atât specialişti, cât şi operatori şi angajarea acestora în cercetare;
- stabilirea mijloacelor materiale necesare pentru desfăşurarea anchetei;
- stabilirea costurilor şi realizarea unui buget al anchetei;
- realizarea calendarului de desfăşurare a activităţilor cercetării prin sondaj;
- stabilirea modalităţilor de verificare şi control a activităţilor.

1.4.1. Asigurarea necesarului uman

Personalul care participă la realizarea unei anchete prin sondaj este compus din mai multe categorii: personal specializat, operatori, personal administrativ.

a. Personalul specializat

În general, o anchetă prin sondaj este alcătuită dintr-o echipă interdisciplinară care are un coordonator ştiinţific. Dimensiunea şi componenţa echipei este dictată de tema şi amploarea anchetei. De exemplu, în cazul unei teme de

cercetare care vizează opinia populației cu privire la activitatea unui primar, echipa specializată poate cuprinde: statisticieni și sociologi.

Statisticienii au în special rolul de a elabora planul de sondaj și de a realiza analiza datelor, însă au o contribuție importantă și la elaborarea modelului de analiză și construirea chestionarului, precum și în elaborarea raportului de cercetare.

Sociologii joacă un rol important în etapa de definire a temei, în etapa de documentare, problematizare, operaționalizare a conceptelor, în formularea obiectivelor și ipotezelor cercetării, în elaborarea chestionarului, în selectarea și instruirea operatorilor. De asemenea, sociologii participă la etapa de analiză a datelor și în realizarea raportului de cercetare.

Personalul specializat implicat în anchetă poate fi personalul angajat permanent al unei instituții sau poate fi și personal angajat temporar în cercetarea respectivă. Selecția personalului cade în sarcina directorului de proiect.

b. Operatorii

Operatorii sunt persoanele angajate temporar în anchetă pentru a realiza etapa de observare sau de culegere a datelor. Funcția acestora constă în administrarea eficientă a chestionarelor, astfel încât erorile de observare să fie cât mai reduse. În general, institutele de sondaj își realizează o rețea de operatori pe teritoriul unei țări sau a unei arii geografice. Această rețea este constituită ca urmare a unui proces de selecție și formare.

Selecția operatorilor se face în funcție de anumite trăsături:

- fizice, care se referă la aspectul plăcut, vârsta tânără, îmbrăcăminte decentă etc.;
- psihice, care au în vedere capacitățile de comunicare ale persoanei, echilibrul și capacitatea de a nu influența persoana care răspunde la întrebări etc.;
- morale, care se referă la corectitudine și onestitate în realizarea sarcinilor acordate.

Formarea operatorilor este activitatea care vizează o bună pregătire a acestora în vederea culegerii datelor. În munca de instruire, operatorii trebuie să-și însușească temeinice cunoștințe despre:

- tema și obiectivele anchetei;
- conținutul chestionarului și forma sa;
- modul de completare a fiecărei întrebări;
- modul de utilizare a informațiilor suplimentare care se oferă în timpul administrării chestionarului;
- consemnele de identificare a persoanelor care participă la anchetă;
- timpul și locul de completare a chestionarului;

- participarea la activitatea de control;
- etc.

c. Personalul administrativ

Această categorie de personal are un rol important în organizarea şi desfăşurarea activităţii de cercetare, asigurând funcţionalitatea echipamentelor, necesarul de materiale şi instrumente etc. Personalul administrativ poate fi compus din:

- managerul proiectului de cercetare sau directorul de proiect, care asigură administrarea fondurilor şi buna desfăşurare a activităţilor;
- administratorul echipamentelor informatice, care asigură funcţionalitatea la parametri normali a reţelelor, a computerelor şi a celorlalte echipamente necesare;
- secretarul / tehnoredactorul, care asigură comunicarea, munca de editare a materialelor necesare anchetei;
- etc.

1.4.2. Asigurarea mijloacelor materiale şi financiare. Bugetul anchetei

Mijloacele materiale necesare într-o anchetă prin sondaj pot fi structurate în următoarele categorii: echipamente informatice, publicaţii, baze de date, consumabile, mijloace materiale auxiliare.

a. Echipamente informatice şi de calcul

În această categorie putem menţiona computerele dotate cu softul necesar, precum şi echipamentele periferice necesare: imprimante, scanere etc. Dacă sondajul este organizat de o firmă de specialitate, această dotare poate să existe în prealabil şi se pot face doar activităţi de upgrade / actualizare, atât pentru echipamentul hard, cât şi pentru soft, sau se pot adăuga şi alte echipamente noi care sunt achiziţionate în etapa de pregătire a sondajului.

b. Cărţi şi diverse publicaţii

Atât activitatea de pregătire a sondajului, cât şi cea de realizare efectivă pot necesita o serie de documente din care se pot desprinde informaţii importante. Pentru documentare se pot utiliza biblioteci şi surse de informaţii care nu implică costuri, dar uneori se impune achiziţionarea unor materiale specifice (cărţi, rapoarte de cercetare din alte studii, reviste etc.)

c. Baze de date

Sondajele statistice, în funcţie de planul de sondaj ales, pot impune utilizarea şi chiar achiziţionarea unor baze de date. Aceste date pot fi disponibile în sistemul statisticii oficiale sau la nivelul unor firme specializate. Frecvent, asemenea baze de date se obţin contra cost, cu o structură şi un volum bine determinate. De exemplu, o bază de sondaj pentru un sondaj aleator la nivelul firmelor de

construcții dintr-o anumită regiune geografică a unei țări nu este disponibilă decât în cadrul sistemului oficial (statistic sau fiscal) sau la nivelul unei firme de specialitate care a reușit să construiască și să întrețină la zi o asemenea bază de date. Pentru un asemenea sondaj, o soluție este cumpărarea unei baze de date care să reprezinte lista unităților deja selectate pentru a constitui eșantionul.

d. Consumabile. Mijloace materiale auxiliare

Pentru anchetele mari, materialele consumabile nu sunt neglijabile (hârtie, mapele operatorilor, elemente de promovare, toner etc.). De asemenea, materialele auxiliare pot fi consistente dacă arealul anchetei este unul semnificativ. În această categorie pot intra echipamentele de comunicare (telefoane, pagere etc.) și mijloacele de transport (mașini) și de cazare. Pentru munca de teren, trebuie să se asigure: comunicarea cu operatorii, pentru a realiza informarea și controlul activității acestora; mijloacele de deplasare a operatorilor, caz în care fie se dispune de mijloacele firmei, fie se închiriază anumite mijloace de transport, fie se asigură căi facile de transport cu mijloacele puse la dispoziție de firmele specializate; mijloacele de cazare și de masă ale operatorilor (îndeosebi atunci când operatorii trebuie să se deplaseze în alte localități).

e. Bugetul

Pe baza necesarului de personal și de materiale, se trece la stabilirea cheltuielilor necesare unei anchete. Instrumentul de bază care asigură delimitarea tipurilor de cheltuieli și, în final, a costului total și a celui unitar este bugetul anchetei.

Pe baza elementelor din buget, se determină următoarele costuri:

- capitole sau tipuri de costuri: costuri cu salariile, cu echipamentele etc.;
- costul total, ca sumă a tuturor costurilor pe capitole sau pe tipuri de costuri;
- costul unitar (c), care se obține împărțind costul total (C) la volumul eșantionului sau la numărul de chestionare administrate (n).

Constrângerile de cost sunt importante în realizarea oricărui sondaj. Cel mai adesea, se cunoaște suma disponibilă pentru realizarea anchetei. Această sumă, ce joacă rolul de cost total, poate determina tipul de sondaj și volumul eșantionului. Dacă nu este impusă o sumă, prioritare vor fi celelalte criterii care decid metoda de sondaj și, în final, volumul eșantionului: precizia dorită, informația disponibilă etc.

Un exemplu de structură de buget pentru o anchetă se prezintă în tabelul 1.3.

Tabelul 1.3. *Bugetul unui sondaj*

Cheltuieli	Unitate	Număr de unităţi	Cost unitar	Cost total
1. Resurse umane				
1.1. Salarii (sume brute)				
1.1.1. Personal administrativ	oră			
1.1.2. Personal specializat (echipe de specialişti)	oră			
1.1.3. Operatori	chestionar			
1.1.4. Consultanţi	oră			
Subtotal Resurse umane				
2. Date şi informaţii				
2.1. Cărţi şi publicaţii	buc.			
2.2. Baze de date				
Subtotal Date şi informaţii				
3. Echipamente şi consumabile				
3.1. Office echipament				
3.2. Soft				
3.3. Consumabile				
3.4. Altele				
Subtotal Echipamente şi bunuri				
4. Costuri aferente proiectului				
4.1. Costuri transport, cazare şi muncă de teren	persoană			
4.2. Costuri utilităţi	lună			
Subtotal Sediu local / costuri aferente proiectului				
5. Alte costuri, servicii				
5.1. Publicaţii	abonament			
5.2. Traduceri	pagină			
5.3. Tipărire chestionare şi alte materiale	pagină			
5.4. Alte costuri				
Subtotal Alte costuri, servicii				
6. Subtotal costuri directe aferente proiectului (1.- 5.)				
7. Costuri administrative 2%				
8. Total costuri eligibile aferente proiectului (6. + 7.)				
9. Costuri neprevăzute 2%				
10. Total costuri (8. + 9.)				

1.4.3. Calendarul cercetării

Fiecare activitate din programul de cercetare are un loc şi un timp de realizare. Cercetarea prin sondaj este un şir sau lanţ de operaţiuni care necesită un timp minim de realizare. Estimarea acestor timpi şi organizarea lor într-un întreg coerent al programului de cercetare se concretizează în calendarul sondajului statistic. Pentru realizarea acestui instrument trebuie luat în considerare că, în general, în practică, timpul estimat cercetării este depăşit cu un procent de până la 150% (Warwick şi Lininger, 1975, p. 35).

Eficienţa sondajului presupune ca timpul de realizare a anchetei să fie cel mai scurt posibil. În general, pentru fiecare activitate se stabileşte un interval de timp care poate oscila între un minim şi un maxim, deoarece pot interveni şi anumite probleme, însă se urmăreşte atingerea limitei inferioare a intervalului. Respectarea termenelor pentru fiecare activitate este deosebit de importantă, deoarece există activităţi care se succed, iar realizarea lor în timpul prevăzut asigură mersul normal al cercetării.

Pentru gestionarea timpului în desfăşurarea activităţilor cercetării este util să se folosească un instrument de management al timpului şi al ordinii activităţilor, aşa cum este diagrama Gantt.

Diagrama Gantt conţine o listă cu toate activităţile ce presupun realizarea sondajului, împreună cu timpul alocat fiecăreia. Unităţile de timp trebuie să fie alese corespunzător, astfel încât să poată fi evaluat cât mai bine timpul alocat fiecărei activităţi. O serie de activităţi se pot realiza în paralel, iar altele presupun o ordine logică. De asemenea, unele activităţi depind de rezultatele activităţilor anterioare. Aşa cum se observă din exemplul prezentat în tabelul 1.4, diagrama Gantt face posibilă delimitarea activităţilor, succesiunea lor, inclusiv realizarea în paralel. Tot în diagramă se precizează volumul de timp necesar pentru fiecare activitate, precum şi timpul total.

Tabelul 1.4. *Model diagramă Gantt pentru un sondaj de opinie*

Nr. crt.	Activitate	L	M	M	J	V	S	L	M	M	J	V	S
1.	Definire problemă. Model de analiză												
2.	Elaborare chestionar												
3.	Plan de sondaj												
4.	Organizare anchetă												
5.	Culegere date												
6.	Analiză date												
7.	Raport cercetare												

În tabelul de mai sus, activitățile sunt prezentate la nivelul cel mai înalt de sinteză, în ipoteza realizării anchetei în două săptămâni, a câte șase zile de lucru fiecare. De regulă, asemenea diagrame se realizează mai detaliat, pe activități specifice la nivelul fiecărei etape a cercetării, pentru a gestiona cât mai eficient timpul.

1.4.4. Organizarea spațiului

În general, anchetele prin sondaj se realizează pe spații mari: pe teritoriul unei țări, al unei zone (zonă de dezvoltare, regiune istorică, județ, zonă rurală sau agricolă), al unui oraș (structurat pe cartiere sau zone, pe străzi etc.). Organizarea anchetei presupune și o atentă organizare a teritoriului în care se desfășoară procesul de culegere a datelor. În acest sens, trebuie rezolvate următoarele probleme: asigurarea informațiilor despre teritoriul anchetei, repartizarea operatorilor și stabilirea itinerariilor.

Instrumentul de lucru pentru această etapă este harta anchetei. Această hartă are la bază harta fizică sau cea turistică a teritoriului pe raza căruia se realizează ancheta. Pe această hartă se detaliază apoi unități de spațiu, itinerarii și puncte de control necesare unei bune desfășurări a etapei de culegere a datelor.

Asigurarea informațiilor despre teritoriu se realizează prin cercetarea de hărți cu infrastructura și clădirile; prin analiza posibilităților de transport (documente informative emise de societățile de transport) și a celor de cazare (hoteluri, pensiuni, case de vacanță etc.).

Repartizarea operatorilor este foarte importantă în special în cazul anchetelor care presupun o mobilitate ridicată. Pentru operatori se stabilesc foarte precis itinerariile pe care trebuie să le urmeze în procesul de culegere a datelor, rutele și mijloacele de transport care se utilizează, locurile și perioada de cazare.

Aceste informații detaliate pot face parte din fișele de operator pentru a facilita procesul de culegere a datelor pe teren. Pentru fiecare operator se fixează un anumit teritoriu în care se realizează ancheta și se oferă toate informațiile necesare pentru a ajunge la unitățile care au fost extrase în eșantion sau care pot participa la anchetă.

1.5. Exerciții

1. Prezentați câteva dintre avantajele și limitele cercetării pe bază de sondaj statistic.

2. Caracteristica esențială a unui sondaj este reprezentativitatea. Prezentați câteva dimensiuni ale conceptului și maniera în care s-a dezvoltat înțelegerea despre reprezentativitatea unui sondaj.

3. Exemplificați etapele cercetării prin sondaj în contextul unui studiu privind comportamentul consumatorului de cafea.

4. Operaționalizați conceptul de comportament al consumatorului de cafea, exemplificând pe cazul unei populații mai omogene, cum ar fi cea a studenților dintr-un oraș universitar.

5. Elaborați două ipoteze în contextul unui studiu privind comportamentul consumatorului de cafea al studenților din orașul Iași.

6. Formulați o întrebare de opinie, una de date factuale și una de motivație pentru un chestionar ce este utilizat într-o anchetă privind comportamentul consumatorului de cafea.

7. În contextul anchetei menționate la exercițiul anterior, precizați ce metodă de elaborare a structurii chestionarului se poate utiliza.

8. Aplicând criteriile de alegere a unei metode de sondaj, argumentați alegerea unui tip de sondaj care respectă criteriul de eficiență în contextul unui studiu privind comportamentul consumatorului de cafea în orașul Iași.

9. Pentru același exemplu de anchetă de la exercițiul anterior, realizați un model de buget, după ce s-a optat pentru un tip de sondaj.

10. În condițiile unei anchete definite la exercițiul 7, să se realizeze o diagramă Gantt pentru a prezenta activitățile cerute de anchetă și timpul necesar desfășurării acestora.

11. În cazul unui sondaj empiric, ce criterii se pot utiliza pentru stabilirea volumului eșantionului?

12. În care dintre situațiile prezentate mai jos se poate organiza un sondaj statistic?

- evaluarea opiniei persoanelor care au emigrat din România despre situația politică existentă în țară;
- studiul violenței domestice în mediul rural din România;
- analiza comportamentului consumatorilor de droguri din București;
- opinia germanilor cu privire la un nou model de autoturism.

13. Ce variante de asigurare a reprezentativității unui eșantion există pentru o anchetă de opinie politică la nivelul populației orașului București?

14. Prezentați câteva modalități de a construi baze de sondaj la nivelul populației regiunii de Nord-Est, luând în calcul posibilități de structurare a persoanelor.

15. Ce presupune etapa de conceptualizare în cadrul demersului unei cercetări științifice?

16. Prezentați câteva modalități de reducere a erorilor de eșantionare în cazul realizării unei anchete la nivelul gospodăriilor din mediul rural din România privind nivelul de sărăcie.

17. Exemplificați și explicați diferența dintre un sondaj statistic și selectarea „la întâmplare" a unui număr de unități dintr-o populație, pentru a culege informații pe o anumită temă.

Capitolul 2

SONDAJUL ALEATOR SIMPLU

Cea mai cunoscută metodă de sondaj probabilist poartă numele de *sondaj aleator simplu*. Sondajul aleator reprezintă metoda de cercetare care are la bază un principiu de selecție stocastic. Extragerea eșantionului reprezintă un eveniment în cadrul unei experiențe aleatoare pentru care se poate realiza calculul probabilității de apariție a unui eșantion, precum și calculul probabilității de includere a unei unități din populație în eșantion. Selecția se poate realiza asigurând aceeași probabilitate de includere în eșantion fiecărei unități din populație (cazul sondajului aleator simplu) sau șanse diferite de includere (cazul sondajului cu probabilități inegale).

În acest capitol ne propunem să oferim cititorului o prezentare a noțiunilor de bază, a procedeelor de selecție și a instrumentelor statistice utilizate în practica sondajelor aleatoare cu probabilități egale. Ne referim, în principal, la prezentarea estimatorilor statistici care se definesc, în acest caz, pentru a fi utilizați în probleme de estimare și testare cu privire la parametrii unei populații.

2.1. Noțiuni și notații

Așa cum am subliniat deja, prin *sondaj aleator simplu* înțelegem metoda de sondaj care are la bază principiul extragerii aleatoare a unităților care compun eșantionul, respectând proprietatea de echiprobabilitate sau condiția de probabilitate egală a fiecărei unități din populație de a fi inclusă în eșantion. Această idee de selecție este echivalentă și cu cea de a conferi aceeași probabilitate unui grup de n unități din populație de a fi extrase sau aceeași probabilitate de extragere pentru fiecare eșantion posibil de obținut din populația de referință.

Înainte de a intra în detaliile acestei metode, în debutul acestui capitol este util să fie prezentate câteva noțiuni și notațiile utilizate pentru individualizarea și identificarea facilă a acestora pe parcursul lucrării.

2.1.1. Populație și eșantion. Parametri și estimații

Populația statistică mai este întâlnită și sub numele de univers sau colectivitate. Reprezintă un agregat sau un ansamblu de unități individuale reale definite cu ajutorul unei caracteristici care determină apartenența fiecărei unități la întreg, având un referent spațial și temporal precis. Populația este de obicei notată cu simbolul P, iar volumul acesteia cu N, reprezentând numărul de unități individuale care compun întregul. Populațiile pot fi structurate în părți, numite sub-populații, cu ajutorul unor criterii utile pentru cercetarea științifică. De exemplu, un studiu la nivelul populației României poate considera sub-populațiile județelor, regiunilor sau mediului rural și urban.

Unitățile care compun populația se numesc unități elementare sau elemente ale populației. Acestea pot fi simple (obiecte, persoane, evenimente etc.) sau complexe (firme, familii etc.) și trebuie să fie bine definite și ușor de identificat la nivelul întregului. Pentru unitatea statistică se mai utilizează și numele de *individ*. În contextul cercetării prin sondaj, se mai poate face distincția între unitatea statistică de observare și unitatea de eșantionare. Prima face trimitere la unitatea care este supusă efectiv procesului de măsurare, iar cea de-a doua este legată de operația de eșantionare. Uneori cele două coincid (de exemplu, un student), iar alteori diferă (de exemplu, unitatea de eșantionare este o firmă, iar cea de observare este managerul).

O cerință esențială a sondajului aleator este aceea de a constitui o *bază de sondaj* (*sampling frame*), adică o listă sau o bază de date cu toate unitățile individuale și cu un element de identificare ce ține cont de tipul și natura populației (pentru o firmă, de exemplu, numele sau codul fiscal, iar pentru o persoană, nume, adresă sau CNP). Baza de sondaj reprezintă instrumentul care permite aplicarea algoritmilor de extragere aleatoare a eșantionului. De asemenea, baza de sondaj poate conține informații suplimentare despre populație, ce pot fi utile în etapa de eșantionare sau în cea de estimare pentru a crește eficiența sondajului.

Cu privire la populațiile statistice interesează studiul unor *caracteristici* sau trăsături care sunt surprinse cantitativ cu ajutorul *variabilelor statistice*. Variabilele reprezintă instrumente care îndeplinesc funcția de măsurare a trăsăturilor analizate. De regulă, acestea se notează cu ajutorul literelor latine majuscule X, Y, Z.

La nivelul populației, pentru o variabilă statistică X se definesc o serie de caracteristici numerice sau *indicatori statistici,* cum ar fi indicatorii tendinței centrale, ai dispersiei sau formei distribuției. Aceste mărimi reale fixe sunt, de regulă, necunoscute și poartă numele de *parametri*. În cercetările pe bază de sondaj, un obiectiv important îl reprezintă estimarea parametrilor cu ajutorul datelor obținute la nivelul eșantionului. Parametrul este definit de valorile variabilei X la nivelul tuturor unităților populației. De obicei, parametrul se notează cu litere grecești (μ, π, σ, λ etc.). De exemplu, dacă notăm generic parametrul cu

θ, atunci se poate scrie ca o funcție de valorile variabilei la nivelul unităților individuale:

$$\theta = f(x_1, x_2, \ldots, x_N)$$

unde f este o funcție măsurabilă.

Prezentăm mai jos câteva exemple de parametri utilizați în cercetările statistice.

– Media populației

Parametrul *medie* se notează cu μ și este definit de relația:

$$E(X) = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

– Valoarea globală (totalul)

În practică, apare frecvent situația de a estima un total, volumul sau valoarea globală (de exemplu, volumul vânzărilor sau al creditelor etc.). Pentru o variabilă X notăm *totalul* prin τ și se obține ca sumă a valorilor individuale ale variabilei la nivelul populației totale:

$$\tau = \sum_{i=1}^{N} x_i$$

Evident,

$$\tau = N \cdot \mu$$

$$\mu = \frac{\tau}{N}$$

– Moment inițial de ordin k

Ca o generalizare a mediei aritmetice, se definește parametrul *moment inițial de ordin k*:

$$\mu_k = \frac{1}{N}\sum_{i=1}^{N} x_i^k$$

Pentru diverse ranguri ($k = 2,3,\ldots$) se obțin parametri ce sunt utili în analiza statistică a distribuției unei populații sau în analiza legăturii dintre variabile.

– Proporția

Considerăm cazul unei variabile alternative (de tip Bernoulli):

$$X: \begin{pmatrix} 0 & 1 \\ 1-\pi & \pi \end{pmatrix}$$

Parametrul π reprezintă *proporţia* sau ponderea unităţilor din populaţie care îndeplinesc o proprietate bine definită A. Rezultă:

$$\pi = \frac{M}{N}$$

unde M este numărul unităţilor din populaţie care îndeplinesc condiţia A.

Ca o observaţie importantă, parametrul proporţie reprezintă media variabilei alternative definite mai sus:

$$E(X) = \pi$$

Cu alte cuvinte, estimarea unei proporţii presupune respectarea unor condiţii similare cu situaţia estimării unei medii.

Pentru o variabilă alternativă, parametrul *total* este:

$$\tau = \sum_{i=1}^{N} x_i = M$$

adică este egal cu numărul unităţilor din populaţie ce respectă proprietatea cerută.

– Varianţa

Varianţa populaţiei pentru o variabilă X se notează cu σ^2 şi este definită de relaţia:

$$V(X) = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

sau

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2 = E(X^2) - [E(X)]^2$$

Pentru o variabilă alternativă, varianţa este de forma:

$$V(X) = \sigma^2 = \pi(1 - \pi)$$

– Abaterea standard

O măsură a gradului de eterogenitate a unei populaţii în raport cu o variabilă X este abaterea medie pătratică sau abaterea standard. Aceasta se obţine ca rădăcină pătrată a varianţei:

$$\sigma = \sqrt{V(X)}$$

– Coeficientul de variație

Acest parametru reprezintă o măsură relativă pentru gradul de omogenitate a populației și se obține ca raport între abaterea standard și media aritmetică:

$$v = \frac{\sigma}{\mu}$$

– Moment centrat de ordin k

O generalizare a variației o reprezintă media abaterilor de ordin k, numit *moment centrat de ordin k*:

$$\sigma^{(k)} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^k$$

Momentele de ordinul 3 și 4 sunt utilizate pentru determinarea parametrilor distribuției unei populații (asimetria și boltirea).

Următorul concept fundamental în cercetarea pe bază de sondaj este *eșantionul*. Este notat cu s (de la *selecție*) și are un volum care este notat cu n, reprezentând numărul de unități extrase din populație. Ca subansamblu de unități al populației, eșantionul se obține în urma procesului de selecție aleatoare care asigură echiprobabilitatea participării la eșantion a tuturor unităților din populație.

Un corespondent al caracteristicilor numerice de la nivelul populației la nivelul eșantionului îl reprezintă *estimațiile* sau *valorile de sondaj*. O estimație se obține la nivelul unui eșantion extras, pe baza datelor culese, și este o funcție a valorilor unei variabile (X) înregistrate prin sondaj:

$$estimația = g(x_1, x_2, \dots, x_n)$$

Estimațiile se notează cu litere latine, iar pentru parametrii definiți mai sus corespund următoarele:

– Media de eșantion

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

– Valoarea globală de eșantion (totalul)

$$T = \sum_{i=1}^{n} x_i$$

$$T = n \cdot \bar{x}$$

– Moment inițial de ordin k de eșantion

$$x^{(k)} = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

– Proporția de eșantion

$$w = \frac{m}{n}$$

unde m reprezintă numărul unităților din eșantion care îndeplinesc proprietatea A;

– Varianța de eșantion:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

În unele pachete program de statistică, prin varianță de eșantion (*sample variance*) se înțelege varianța modificată, calculată prin relația:

$$s'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Pentru o variabilă alternativă, varianța la nivel de eșantion este:

$$s^2 = w(1-w)$$

– Abaterea standard de eșantion

$$s = \sqrt{s^2}$$

– Coeficientul de variație

$$v = \frac{s}{\bar{x}}$$

– Moment centrat de ordin k de eșantion

$$s^{(k)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

Estimațiile sunt mărimi reale și cunoscute, caracteristici numerice calculate pe baza datelor de sondaj și stau la baza procesului de estimare a parametrilor.

2.1.2. Procedee de selecție

Procedeul de selecție presupune o experiență aleatoare ce oferă posibilitatea de a calcula probabilitatea de includere a fiecărei unități individuale a populației în eșantion și probabilitatea de a extrage un eșantion din mulțimea de eșantioane

posibil de extras. Selecția aleatoare simplă, care asigură aceeași probabilitate de includere în eșantion pentru fiecare unitate din populație, se poate realiza cel puțin în trei moduri: selecție cu revenire, fără revenire și sistematică.

a. Selecția aleatoare cu revenire

Are la bază schema aleatoare a urnei cu bilă revenită. Din populație este extrasă o unitate, se înregistrează rezultatul, apoi unitatea este reintrodusă în baza de sondaj și urmează o nouă extragere. În acest mod se asigură atât independența extragerilor, cât și echiprobabilitatea lor. Fiecare unitate a populației are aceeași șansă de a fi inclusă în eșantion, probabilitate ce se poate calcula după regula clasică.

Numărul de eșantioane care se pot extrage cu revenire este:

$$K = N^n$$

În extragerea aleatoare, se poate calcula *probabilitatea de extragere a unui eșantion* folosind regula clasică. Pentru selecția cu revenire, această probabilitate este:

$$p(s) = \frac{1}{K} = \frac{1}{N^n}$$

unde

$$p(s) \geq 0, \sum_{s=1}^{k} p(s) = 1$$

Probabilitatea de includere a unei unități în eșantion poartă și numele de probabilitate de includere de ordinul întâi și este definită astfel:

$$p_i = P(i \in s), i = \overline{1,N}$$

Pentru studiul probabilităților, se poate defini variabila:

$$I_{i\in s} = \begin{cases} 1, \text{pt. } i \in s, i = \overline{1,N} \\ 0, \text{în rest} \end{cases}$$

care admite proprietățile:

- $E(I_{i\in s}) = p_i$;
- $V(I_{i\in s}) = p_i(1 - p_i)$;
- $\sum_{i=1}^{N} I_{i\in s} = n \Rightarrow \sum_{i=1}^{N} p_i = n.$

Deoarece suma probabilităților de includere nu este egală cu 1, acestea nu pot constitui o distribuție probabilistă.

Se poate însă construi distribuția de probabilitate de forma:

$$\left(\frac{p_1}{n}, \frac{p_2}{n}, \ldots, \frac{p_N}{n}\right)$$

$$\sum_{i=1}^{N} \frac{p_i}{n} = 1$$

Din cele de mai sus, deoarece probabilitățile de includere sunt egale, rezultă:

$$p_i = p = \frac{n}{N}$$

Mărimea n/N poartă numele de rată sau *fracție de sondaj*.

Probabilitatea de includere se poate calcula luând în considerare specificul extragerii aleatoare cu revenire. Probabilitatea de a extrage o unitate din populație la prima extragere este egală cu $1/N$. Deoarece aceeași unitate poate intra în eșantion și la a doua extragere, dacă urna nu se modifică, probabilitatea nu se modifică. Rezultă că, în total, în cele n extrageri, probabilitatea ca o unitate anume să participe la eșantion este:

$$p = \frac{1}{N} + \frac{1}{N} + \cdots + \frac{1}{N} = \sum_{i=1}^{n} \frac{1}{N} = \frac{n}{N}$$

Extragerea cu revenire are îndeosebi o valoare teoretică ce permite construirea instrumentelor statistice necesare pentru procesul de estimare. În practică, se preferă extragerea fără revenire deoarece nu are sens extragerea unui eșantion care să conțină o unitate de mai multe ori. La limită, dacă populația de referință are un volum foarte mare, se poate accepta acest procedeu de selecție ca unul de referință.

Exemplu

Dintr-o populație de studenți a unei universități ($N = 30000$) se extrage aleator simplu cu revenire un eșantion de volum $n = 300$.

Numărul de eșantioane posibil de extras este $K = 30000^{300}$, iar probabilitatea de extragere a unei unități în eșantion este:

$$p = \frac{n}{N} = \frac{300}{30000} = \frac{1}{100}$$

b. Extragerea aleatoare fără revenire

În acest caz, se apelează la schema probabilistă a urnei cu bilă fără revenire. Adică, după ce o unitate este extrasă din urnă, este înregistrată și rămâne în afara urnei. Procedeul continuă până la extragerea tuturor celor n unități din populație,

respectând aceeași schemă. În acest caz, numărul posibil de eșantioane care pot fi extrase este dat de relația:

$$K = C_N^n$$

Pentru selecția fără revenire, probabilitatea de extragere a unui eșantion este:

$$p(s) = \frac{1}{K} = \frac{1}{C_N^n}$$

unde

$$p(s) \geq 0, \sum_{s=1}^{k} p(s) = 1$$

În acest caz, pentru probabilitățile de includere, pe lângă cele de ordinul întâi, sunt importante probabilitățile de ordinul doi:

$$p_{ij} = P(i \in s \cap j \in s), i \neq j, i, j = \overline{1, N}$$

Pentru probabilitățile de includere de ordinul doi se pot defini variabilele indicatoare:

$$I_{i\in s, j\in s} = \begin{cases} 1, \text{pt. } i \in s, j \in s, i \neq j, i, j = \overline{1, N} \\ 0, \text{în rest} \end{cases}$$

unde

$$\sum_{j=1}^{N} I_{i\in s, j\in s} = (n-1) I_{i\in s} \Rightarrow \sum_{j=1}^{N} p_{ij} = (n-1) p_i$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} I_{i\in s, j\in s} = n(n-1) \Rightarrow \sum_{i=1}^{N}\sum_{j=1}^{N} p_{ij} = n(n-1)$$

Pentru acest caz se poate demonstra că probabilitatea de ordinul întâi rămâne constantă pe tot parcursul extragerii, adică se păstrează condiția de echi-probabilitate. De exemplu, aplicând regula clasică, probabilitatea ca o unitate din populație să intre în eșantion este dată de relația:

$$p_i = p = \frac{C_{N-1}^{n-1}}{C_N^n} = \frac{n}{N}$$

În schimb, probabilitatea de ordinul doi se poate calcula astfel:

$$p_{ij} = \frac{n(n-1)}{N(N-1)}$$

Exemplu

Dintr-o populație de studenți a unei universități ($N = 30000$) se extrage aleator simplu fără revenire un eșantion de volum $n = 300$.

Numărul de eșantioane posibil de extras este $K = C_{30000}^{300}$, iar probabilitatea de extragere a unei unități în eșantion este:

$$p = \frac{n}{N} = \frac{300}{30000} = \frac{1}{100}$$

c. Selecția sistematică

În practică, există situații în care este dificil de obținut o bază de sondaj (un exemplu tipic este cazul sondajelor din ziua alegerilor) și, prin urmare, de aplicat un algoritm de extragere aleator simplu așa cum am arătat mai sus. Dacă la nivelul populației unitățile statistice pot fi identificate ușor în procesul de observare și se consideră că ordinea acestora la nivelul populației este una aleatoare (de exemplu, venirea la vot a cetățenilor de pe listele electorale), atunci se poate realiza selecția după un procedeu simplu care poartă numele de extragere cu pas de numărare sau sistematică.

Dacă se cunoaște volumul eșantionului (n) și volumul populației (N), ideea este să se calculeze un pas de numărare:

$$PAS = \frac{N}{n}$$

Pasul reprezintă instrumentul de selecție și este distanța dintre două unități din populație care sunt incluse în eșantion cu ajutorul acestei metode. Extragerea eșantionului presupune alegerea în mod aleatoriu (cu ajutorul unui algoritm de selecție) a unei prime unități de extragere din populație, iar apoi se aplică pasul de numărare pentru identificarea celorlalte $n - 1$ unități. Utilizarea pasului asigură parcurgerea întregii populații și se evită selecțiile multiple.

Metoda dă rezultate bune și în cazul în care nu se cunoaște exact volumul populației (de exemplu, cazul clienților unui magazin), ci doar se poate face o apreciere. În schimb, cunoscând volumul eșantionului, se poate fixa un pas de numărare care să asigure selecția numărului dorit de unități.

Exemplu

Dintr-o populație de studenți a unei universități ($N = 30000$) se extrage aleator simplu un eșantion de volum $n = 300$ după metoda pasului de numărare. Conform metodei, pasul de numărare este:

$$PAS = \frac{30000}{300} = 100$$

Dacă se extrage aleator o primă unitate din primele 100 de unități, de exemplu unitatea 65, atunci următoarea unitate selectată în eșantion este 165 etc.

2.1.3. Selecții aleatoare. Estimatori. Distribuții de selecție

Pentru înțelegerea detaliilor care privesc eșantionarea aleatoare simplă, considerăm cazul selecțiilor aleatoare cu revenire. Modificările care apar în cazul selecției fără revenire vor fi precizate punctual în capitolele următoare, în discuția legată de estimarea parametrilor unei populații. Prezentăm mai jos, punctual, principalele instrumente care se construiesc în contextul selecției aleatoare cu revenire.

a. Variabile aleatoare de selecție

Din perspectiva teoriei selecției (un capitol al statisticii matematice), extragerea aleatoare a unui eșantion reprezintă o experiență aleatoare care presupune o mulțime de evenimente elementare (notată de regulă cu Ω). Numărul total de eșantioane posibil de extras este:

$$K = N^n$$

Pentru o variabilă X, un eșantion posibil de extras reprezintă un set de n valori dintre valorile posibile ale variabilei care rezultă în urma selecției. Cele K eșantioane aleatoare posibil de extras pot fi analizate împreună ca n realizări structurate astfel:

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{K1} & x_{K2} & \dots & x_{Kn} \end{pmatrix}$$

$$(X_1 \;\; X_2 \;\; \dots \;\; X_n)$$

Fiecare linie din structura matriceală de mai sus reprezintă un eșantion posibil sau valorile posibile ale variabilei X organizate sub formă de eșantioane posibil de extras. Dacă pentru fiecare coloană din matrice se definește o variabilă aleatoare (care are ca valori posibile valorile lui X pentru prima unitate din toate eșantioanele posibil de extras), rezultă o structura matematică de tip vector aleator care se definește astfel:

$$V: (X_1, X_2, \dots, X_n)$$

unde variabilele X_i se numesc *variabile aleatoare de selecție.*

Se poate demonstra că variabilele aleatoare de selecție construite prin schema probabilistă prezentată mai sus sunt identic repartizate cu variabila X (au aceeași lege de probabilitate, deci aceeași parametri ai distribuției) și sunt independente probabilistic.

Aşadar:

$$E(X_i) = E(X)$$

$$V(X_i) = V(X)$$

$$i = \overline{1,n}$$

Rezultă că un eşantion poate fi privit ca o selecţie de volum n, adică o realizare posibilă a unei experienţe aleatoare sau o valoare posibilă a *vectorului aleator* V:

$$v: (x_1, x_2, \ldots, x_n)$$

Exemplu

Dintr-o populaţie compusă din numerele naturale de la 1 la 6 se extrag aleator un număr de 36 de eşantioane formate din 2 numere, adică $K = 6^2$ eşantioane posibil de extras. Aceste eşantioane sunt prezentate în tabelul de mai jos.

(1, 1)	(2, 1)	(3, 1)	(4, 1)	(5, 1)	(6, 1)
(1, 2)	(2, 2)	(3, 2)	(4, 2)	(5, 2)	(6, 2)
(1, 3)	(2, 3)	(3, 3)	(4, 3)	(5, 3)	(6, 3)
(1, 4)	(2, 4)	(3, 4)	(4, 4)	(5, 4)	(6, 4)
(1, 5)	(2, 5)	(3, 5)	(4, 5)	(5, 5)	(6, 5)
(1, 6)	(2, 6)	(3, 6)	(4, 6)	(5, 6)	(6, 6)

b. Statistici

O *statistică* este o variabilă aleatoare obţinută ca o combinaţie a variabilelor aleatoare de selecţie X_i. Dacă notăm o statistică cu S, aceasta se obţine cu ajutorul unei funcţii numărabile g care compune elementele vectorului V:

$$S = g \circ V$$

De exemplu, dacă aplicaţia g este o medie, se obţine statistica:

$$S = \frac{1}{n}\sum_{i=1}^{n} X_i$$

care este o variabilă aleatoare ale cărei valori posibile sunt mediile tuturor eşantioanelor de volum n posibil de extras din populaţia N.

O asemenea valoare posibilă a statisticii S este media de eşantion:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

c. Distribuția de selecție

Statisticile sunt variabile aleatoare construite cu ajutorul selecțiilor de volum n sau a eșantioanelor posibil de extras dintr-o populație respectând principiul echiprobabilității. Ceea ce interesează cu privire la aceste variabile aleatoare este legea de probabilitate. Distribuția probabilistică a unei statistici se numește *distribuție de selecție.*

Pentru statisticile mai frecvent utilizate în analiza statistică se determină atât legile de distribuție, cât și proprietățile acestora. Pentru eșantioane de volum mare (așa cum sunt cele obținute pentru sondajul aleator simplu), unele proprietăți sunt acceptate ca proprietăți limită (asimptotice).

d. Estimatorii statistici

Pentru fiecare parametru de estimat se identifică o statistică ce îndeplinește o serie de condiții de calitate. *Estimatorul* este o statistică aleasă sau construită convenabil pentru fiecare parametru al populației. Estimatorul se notează cu litere grecești care corespund parametrilor, având însă deasupra simbolul " ^ ".

De exemplu, pentru parametrul θ, notăm estimatorul cu $\hat{\theta}$. Acesta reprezintă o combinație a variabilelor aleatoare de selecție, ca și în cazul statisticilor:

$$\hat{\theta} = g(X_1, X_2, \ldots, X_n)$$

Funcția g se alege în procesul de identificare a celui mai bun estimator, adică a statisticii care respectă un set de proprietăți privind calitatea rezultatelor.

Estimatorul este instrumentul principal care permite estimarea unui parametru. Calitatea procesului depinde de proprietățile estimatorilor (care vor fi prezentate mai jos).

e. Erori de estimare

De regulă, există o diferență între o estimație și parametrul pe care îl estimează, ceea ce reprezintă o *eroare de estimare.* Această eroare poate fi evaluată cu ajutorul proprietăților estimatorilor.

Pentru definirea proprietăților estimatorilor, este important să se pornească de la o măsură a erorii de estimare care se numește *eroarea medie pătratică* (o medie a abaterii valorilor estimatorului de la parametru):

$$E(\hat{\theta} - \theta)^2 = V(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2$$

În relația de mai sus, expresia

$$V(\hat{\theta}) = E\left(\hat{\theta} - E(\hat{\theta})\right)^2$$

este *varianța estimatorului* (o măsură a gradului de împrăștiere a valorilor sale posibile față de media lor sau față de media estimatorului), iar expresia

$$B(\theta) = E(\hat{\theta}) - \theta$$

reprezintă *deplasarea* sau *biais*-ul estimatorului.

O bună estimare statistică impune ca eroarea de estimare de mai sus să fie minimă (ea este practic zero dacă estimația este identică cu parametrul), adică o varianță minimă și un biais minim.

O altă măsură a erorii de estimare se numește *eroare maxim admisibilă* sau *eroare limită*. Eroarea limită este acea eroare care are la bază legea de distribuție a estimatorului și reprezintă eroarea maxim admisibilă a unui estimator:

$$\Delta_{\hat{\theta}} = t \cdot \sigma_{\hat{\theta}}$$

unde:

- t reprezintă un coeficient de probabilitate care se determină în funcție de distribuția de selecție a estimatorului și probabilitatea cu care se dorește garantarea rezultatelor (această probabilitate este fixată de cercetător);
- $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$ reprezintă abaterea medie pătratică a estimatorului.

De obicei, în expresia abaterii medii pătratice, apar parametrii distribuției de selecție a estimatorului, care depind de parametrii populației totale. Cum acești parametri nu sunt cunoscuți, intervine problema estimării erorii limite, adică estimarea preciziei cu care se garantează rezultatele.

În practică, această eroare limită se fixează de către cercetător pentru a calcula apoi volumul eșantionului care să permită estimarea parametrilor în aceste condiții de eroare maxim admisă.

f. Proprietățile estimatorilor

Estimarea statistică se realizează în baza următoarelor proprietăți mai importante ale estimatorilor: nedeplasare, convergență în probabilitate, eficiență, convergență în repartiție.

Nedeplasarea

Un estimator se numește *nedeplasat* sau bine centrat (fără biais) dacă media sa este egală cu parametrul sau deplasarea este nulă:

$$E(\hat{\theta}) = \theta \text{ sau } B(\theta) = 0$$

Proprietatea de nedeplasare presupune ca estimatorul să aibă ca tendință centrală parametrul și nu o altă valoare. A respecta această condiție înseamnă a spune că parametrul reprezintă speranța matematică a estimatorului (o valoare cu probabilitate de apariție mare - cea mai mare dacă distribuția este simetrică).

Există situaţii când estimatorul este deplasat, dar deplasarea devine neglijabilă la un volum al eşantionului suficient de mare (vorbim despre un estimator asimptotic nedeplasat). Deplasarea depinde de parametru, ca atare este necunoscută. Deoarece parametrul este estimat pe baza datelor de la nivelul unui eşantion, rezultă că şi deplasarea este estimată, valoarea sa reală fiind necunoscută.

Nedeplasarea poate fi privită şi ca o proprietate asimptotică. Astfel, estimatorul $\hat{\theta}$ se numeşte *asimptotic nedeplasat* pentru parametrul θ dacă:

$$lim_{n\to\infty} E(\hat{\theta}) = \theta \text{ sau } lim_{n\to\infty} B(n,\theta) = 0$$

Convergenţa în probabilitate

Această proprietate impune o condiţie de volum a eşantionului: dacă acesta este suficient de mare, atunci orice valoare posibilă a estimatorului (orice estimaţie) converge către parametru. Această proprietate se poate scrie astfel:

$$(\hat{\theta}_n)_{n\in N} \xrightarrow{p} \theta$$

sau

$$P(\forall \varepsilon > 0,\ \exists n_\varepsilon,\ n > n_\varepsilon,\ |\hat{\theta}_n - \theta| < \varepsilon) = 1$$

Proprietatea de convergenţă prezentată mai sus este o variantă a celebrei legi statistice numite *legea numerelor mari* şi mai poartă numele de *consistenţă slabă.*

Sondajele aleatoare impun o condiţie de volum care este cerută de această proprietate. În practică, acest lucru se traduce în necesitatea de a construi eşantioane mari pentru a asigura estimaţii de calitate.

Observaţie

Se poate demonstra uşor că, pentru un estimator $\hat{\theta}$, dacă sunt îndeplinite următoarele condiţii (condiţii suficiente pentru convergenţa în probabilitate):

$$\begin{cases} \lim\limits_{n\to\infty} E(\hat{\theta}) = \theta \\ \lim\limits_{n\to\infty} V(\hat{\theta}) = 0 \end{cases}$$

atunci estimatorul este convergent.

Eficienţa

Există situaţii când pentru un parametru se pot obţine mai mulţi estimatori (de exemplu, pentru o distribuţie Poisson). În aceste condiţii, apare problema alegerii celui „mai bun" estimator, deci a determinării unui criteriu de alegere a celui mai bun estimator.

Conform inegalităţii lui Cebîşev, pentru un estimator $\hat{\theta}$ care admite varianţă are loc proprietatea:

$$P\left[\left(\hat{\theta}-\theta\right)\geq\varepsilon\right]\leq\frac{V\left(\hat{\theta}\right)}{\varepsilon^2},(\forall)\,\varepsilon>0$$

Din această relație rezultă că un criteriu de alegere a estimatorilor este minimizarea variației acestora.

Dintre doi estimatori posibili pentru un parametru, estimatorul cu varianța cea mai mică se numește un estimator mai eficient decât celălalt (*eficiență relativă*). Dintre doi estimatori posibili pentru un parametru, este preferabil estimatorul cu varianța mai mică. Estimatorul cu varianța minimă este un *estimator eficient*.

Pentru un estimator eficient se respectă condiția:

$$V\left(\hat{\theta}\right)=E\left(\hat{\theta}-E\left(\hat{\theta}\right)\right)^2=min$$

O modalitate de evaluare a acestei condiții de minim o oferă inegalitatea Rao-Cramer. În această inegalitate apare varianța estimatorului și cantitatea de informație pe care o aduce un eșantion aleator extras dintr-o populație pentru parametrul θ, notată cu $I_n(\theta)$. Inegalitatea este dată prin relația:

$$V\left(\hat{\theta}\right)\geq\frac{1}{I_n(\theta)}$$

Dacă f este funcția de probabilitate a variabilei X, este derivabilă de două ori în raport cu parametrul, iar X_i sunt variabilele aleatoare de selecție, atunci cantitatea de informație se obține prin relația:

$$I_n(\theta)=E\left[\frac{\partial}{\partial\theta}lnf(x_1,x_2,\ldots.,x_n;\theta)\right]^2=-E\left[\frac{\partial^2}{\partial\theta^2}lnf(x_1,x_2,\ldots.,x_n;\theta)\right]$$

$$f(x_1,x_2,\ldots.,x_n;\theta)=\prod_{i=1}^{n}f(x_i;\theta)$$

poartă numele de *funcție de verosimilitate* a parametrului θ.

Pentru un sondaj aleator, informația obținută pentru un parametru θ la o realizare a unui eșantion este de n ori informația obținută de o singură observație:

$$I_n(\theta)=nI(\theta)$$

$$I(\theta)=-E\left[\frac{\partial^2}{\partial\theta^2}lnf(X;\theta)\right]$$

Condiția de eficiență a unui estimator presupune condiția de egalitate din relația Rao-Cramer. Atunci când varianța estimatorului respectă relația:

$$V(\hat{\theta}) = \frac{1}{I_n(\theta)}$$

estimatorul $\hat{\theta}$ se numeşte eficient.

Convergenţa în repartiţie (teorema limită centrală)

Această proprietate impune o condiţie de volum pentru combinația liniară a variabilelor aleatoare de selecție X_i, notată cu S_n:

$$S_n = \sum_{i=1}^{n} X_i$$

Prin standardizare se obține:

$$\tilde{S}_n = \frac{S_n - E(S_n)}{\sqrt{V(S_n)}}$$

Dacă volumul eşantionului creşte peste o anumită limită, atunci variabila aleatoare obţinută prin standardizare urmează o lege de repartiţie normală standard:

$$(\tilde{S}_n)_{n\in N} \xrightarrow{rep} Z \sim N(0,1)$$

Proprietatea se aplică cu uşurinţă tuturor estimatorilor care se obţin ca o combinaţie liniară a variabilelor aleatoare de selecţie.

2.2. Statistici utilizate în cazul sondajului aleator simplu

În acest subcapitol sunt prezentate o serie de statistici mai frecvent utilizate în cercetările pe bază de sondaj: media şi totalul de selecţie, proporţia de selecţie, varianţa de selecţie.

2.2.1. Media de selecţie

Pentru fiecare statistică se vor prezenta definiţia şi o serie de proprietăţi utile în analiza statistică.

Definiţie

Se numeşte medie de selecţie statistica definită prin relaţia:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Valorile posibile ale acestei variabile aleatoare sunt mediile de sondaj ale tuturor eşantioanelor de volum n posibil de extras dintr-o populaţie.

Pentru un eşantion extras v: $(x_1, x_2, \ldots, x_n)$, media de selecţie ia o valoare care este media aritmetică de eşantion şi reprezintă estimaţia parametrului medie:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Dacă luăm în calcul mulțimea eșantioanelor posibil de extras, media de selecție se poate prezenta ca o variabilă aleatoare ce admite câte o valoare pentru fiecare eșantion posibil de extras:

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{K1} & x_{K2} & \dots & x_{Kn} \end{pmatrix} \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \dots \\ \bar{x}_K \end{pmatrix}$$

sau

$$\hat{\mu}: (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_K)$$

Exemplu

Pentru cele 36 de eșantioane prezentate anterior, media de selecție presupune variabila aleatoare cu următoarele valori posibile:

1,0	1,5	2,0	2,5	3,0	3,5
1,5	2,0	2,5	3,0	3,5	4,0
2,0	2,5	3,0	3,5	4,0	4,5
2,5	3,0	3,5	4,0	4,5	5,0
3,0	3,5	4,0	4,5	5,0	5,5
3,5	4,0	4,5	5,0	5,5	6,0

Proprietăți

Media de selecție admite o serie de proprietăți, dintre care prezentăm câteva mai importante.

Media de selecție este un estimator nedeplasat pentru parametrul medie

$$E(\hat{\mu}) = \mu$$

Demonstrația este relativ simplă și se bazează pe proprietățile variabilelor aleatoare de selecție X_i $(i = \overline{1,n})$ - sunt identic repartizate cu variabila X și sunt independente probabilistic. Rezultă că:

$$E(X_i) = E(X) = \mu$$

$$V(X_i) = V(X) = \sigma^2$$

În concluzie,

$$E(\hat{\mu}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n \cdot \mu}{n} = \mu$$

Media mediei de selecție se poate calcula și cu ajutorul probabilității de includere de ordinul întâi, pe baza variabilelor aleatoare care indică includerea în eșantion a unei unități.

Putem scrie relațiile:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{N} x_i I_{i\in s}$$

$$E(\hat{\mu}) = E\left(\frac{1}{n}\sum_{i=1}^{N} x_i I_{i\in s}\right) = \frac{1}{n}\sum_{i=1}^{N} x_i E(I_{i\in s}) = \frac{1}{n}\sum_{i=1}^{N} x_i \frac{n}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i = \mu$$

Exemplu

Pentru cele 36 de eșantioane prezentate anterior, media mediei de selecție este:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1+2+3+4+5+6}{6} = 3{,}5$$

$$E(\hat{\mu}) = \frac{1}{K}\sum_{j=1}^{K} \bar{x}_j = \frac{126}{36} = 3{,}5$$

Deci

$$E(\hat{\mu}) = \mu$$

Media de selecție este un estimator convergent pentru parametrul medie

Conform teoremei lui Hincin, pentru variabile aleatoare de selecție X_i are loc proprietatea:

$$P(\forall \varepsilon > 0, \exists n_\varepsilon, n > n_\varepsilon, |\hat{\mu} - \mu| \geq \varepsilon) = 0$$

sau

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} E(X_i) = \mu$$

Cu alte cuvinte, dacă volumul eșantionului este suficient de mare, orice valoare a estimatorului (orice medie de eșantion) se află în vecinătatea parametrului. Această proprietate este esențială în cazul estimării mediei populației. Deoarece eșantioanele aleatoare respectă această condiție de volum, proprietatea de mai sus se consideră îndeplinită.

Media de selecție este un estimator eficient pentru parametrul medie

Varianța mediei de selecție se obține distinct în funcție de tipul de extragere, ceea ce implică utilizarea probabilităților de includere de ordinul întâi și doi. Pentru ambele situații însă, se poate demonstra că media de selecție este un estimator eficient.

Notăm varianța mediei de selecție astfel:

$$V(\hat{\mu}) = \sigma_{\hat{\mu}}^2$$

Pentru extragerea cu revenire

$$\sigma_{\hat{\mu}}^2 = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Să arătăm că estimatorul $\hat{\mu}$ este eficient, considerând că variabila X urmează o lege de distribuție normală: $X \sim N(\mu, \sigma^2)$. Pentru această distribuție au loc relațiile:

$$lnf(x,\mu) = ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) = -ln\sigma\sqrt{2\pi} - \frac{(x-\mu)^2}{2\sigma^2}$$

$$\frac{\partial}{\partial\mu} lnf(x,\mu) = \frac{(x-\mu)}{\sigma^2}$$

$$\frac{\partial^2}{\partial\mu^2} lnf(x,\mu) = -\frac{1}{\sigma^2}$$

Rezultă:

$$I(\mu) = \frac{1}{\sigma^2}$$

iar

$$I_n(\mu) = \frac{n}{\sigma^2}$$

Concluzia imediată este aceea că

$$V(\hat{\mu}) = \frac{1}{I_n(\mu)}$$

adică se respectă condiția de egalitate din relația Rao-Cramer, deci estimatorul este eficient.

Pentru extragerea fără revenire

În acest caz, se poate demonstra că varianța mediei de selecție este:

$$V(\hat{\mu}) = \frac{N-n}{n(N-1)}\sigma^2$$

Într-adevăr, varianța se poate calcula cu ajutorul variabilelor aleatoare indicator:

$$V(\hat{\mu}) = E(\hat{\mu} - \mu)^2$$

$$V(\hat{\mu}) = E\left(\frac{1}{n^2}\sum_{i=1}^{N}(x_i-\mu)^2 I_{i\epsilon s}^2 + \frac{1}{n^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i-\mu)(x_j-\mu)I_{i\epsilon s, j\epsilon s}\right)$$

Deoarece:

$$E(I_{i\epsilon s}^2) = E(I_{i\epsilon s}) = p_i = \frac{n}{N}$$

$$E(I_{i\epsilon s, j\epsilon s}) = p_{ij} = \frac{n(n-1)}{N(N-1)}$$

iar

$$\left(\sum_{i=1}^{N}(x_i-\mu)\right)^2 = \sum_{i=1}^{N}(x_i-\mu)^2 + \sum_{i=1}^{N}\sum_{j=1}^{N}(x_i-\mu)(x_j-\mu) = 0$$

rezultă că

$$\sigma_{\hat{\mu}}^2 = \frac{\sigma^2}{n}\cdot\frac{N-n}{(N-1)}$$

Dacă volumul populației este suficient de mare, atunci se poate aprecia că diferența dintre N și $N-1$, din relația de mai sus, este neglijabilă și se poate rescrie relația astfel:

$$V(\hat{\mu}) = \frac{\sigma^2}{n}\cdot\frac{N-n}{N} = \frac{\sigma^2}{n}(1-f)$$

unde $f = n/N$ și se numește *fracție de sondaj* sau *factor de corecție* pentru populații finite. Acest factor corectează varianța estimatorului mediei din cazul extragerii cu revenire (caz în care populația este considerată foarte mare sau tinde la infinit).

Rezultă:

$$\sigma_{\hat{\mu}}^2 = \frac{\sigma^2}{n}(1-f)$$

Observație

Între cele două tipuri de extragere se poate observa ușor că există o diferență de precizie care rezultă din compararea variațelor pentru estimatorul medie. Această diferență apare datorită coeficientului $(1-f)$. Dacă fracția de sondaj sau raportul de mărime dintre eșantion și populație este semnificativ, atunci și diferența dintre cele două variațe este semnificativă. În acest caz, sondajul fără revenire este mai precis, pentru că varianța estimatorului este mai mică. Invers, dacă fracția de sondaj este nesemnificativă (volumul populației este foarte mare în comparație cu eșantionul sau tinde la infinit), atunci între cele două tipuri de extragere nu există o diferență semnificativă în materie de precizie.

Convergența în repartiție a mediei de selecție

Aplicată estimatorului medie de selecție, teorema limită centrală ne spune că, indiferent care este legea de repartiție a variabilei X, dacă volumul eșantionului este suficient de mare, atunci estimatorul standardizat urmează o lege de repartiție normală standard:

$$\left(\frac{\hat{\mu}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)_{n\in N} \xrightarrow{rep} Z\sim N(0,1)$$

Această proprietate este esențială pentru că oferă o lege de repartiție pentru estimator în condițiile existenței unui eșantion de volum suficient de mare.

Exemplu

Pentru cele 36 de eșantioane prezentate anterior, distribuția mediei de selecție este prezentată mai jos. Se observă că această distribuție a mediilor eșantioanelor posibil de extras din populația de referință se apropie de forma distribuției normale, în ciuda faptului că volumul eșantionului este mic.

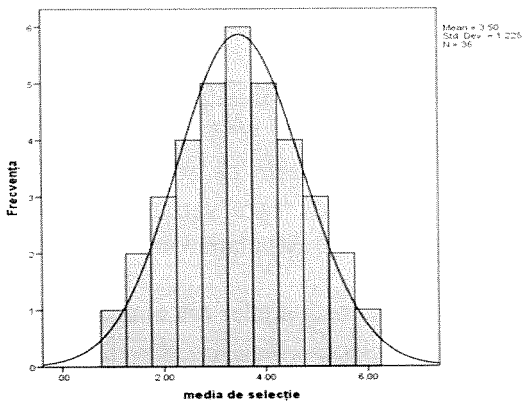


Figura 2.1. *Distribuția mediei de selecție*
Sursa: Realizat de autor. Date convenționale

2.2.2. Valoarea globală sau totalul de selecție

Definiție

Pentru o variabilă X notăm totalul prin:

$$\tau = \sum_{i=1}^{N} x_i$$

$$\tau = N \cdot \mu$$

Un estimator nedeplasat pentru valoarea globală este:

$$\hat{\tau} = N \cdot \hat{\mu} = \frac{N}{n} \sum_{i=1}^{n} X_i$$

unde:

$$E(\hat{\tau}) = N \cdot E(\hat{\mu}) = N \cdot \mu = \tau$$

La nivel de eșantion, o estimație a totalului este o valoare realizată a estimatorului:

$$T = N \cdot \bar{x} = \frac{N}{n} \sum_{i=1}^{n} x_i$$

Proprietăți

Pentru acest estimator se poate determina ușor varianța pentru cele două tipuri de selecție.

Cazul extragerii cu revenire

$$V(\hat{\tau}) = \sigma_{\hat{\tau}}^2 = N^2 \cdot V(\hat{\mu}) = N^2 \cdot \frac{\sigma^2}{n}$$

Cazul extragerii fără revenire

$$V(\hat{\tau}) = \sigma_{\hat{\tau}}^2 = N^2 \cdot V(\hat{\mu}) = N^2 \cdot \frac{N-n}{N} \cdot \frac{\sigma^2}{n} = N(N-n)\frac{\sigma^2}{n}$$

Un estimator al acestei varianțe este:

$$\hat{\sigma}_{\hat{\tau}}^2 = N(N-n) \cdot \frac{\hat{\sigma}'^2}{n}$$

Estimația la nivelul unui eșantion este de forma:

$$s_{\hat{\tau}}^2 = N(N-n) \cdot \frac{s'^2}{n}$$

2.2.3. Moment inițial de selecție de ordin k

Definiție

Pentru parametrul moment inițial de ordin k se definește estimatorul:

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

unde X_i sunt variabilele aleatoare de selecție, iar $k \in N^*$.

Proprietăți

Se poate demonstra că acest estimator are următoarele proprietăți:

- este nedeplasat:

$$E(\hat{\mu}_k) = \mu_k$$

- are varianța și momentul de ordinul 2:

$$V(\hat{\mu}_k) = \frac{\mu_{2k} - \mu_k}{n}$$

unde:

$$\mu_{2k} = E(X^{2k})$$

- este convergent:

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right)_{n\in N} \xrightarrow{p} \mu_k$$

2.2.4. Proporția de selecție

Parametrul proporție poate fi tratat ca o medie a unei variabile binare. Cu alte cuvinte, toate elementele prezentate pentru parametrul medie pot fi utilizate și în cazul definirii unui estimator pentru proporție. În această logică, prezentăm definiția și proprietățile estimatorului *proporția de selecție*.

Definiție

Se consideră variabila alternativă:

$$X: \begin{pmatrix} 0 & 1 \\ 1-\pi & \pi \end{pmatrix}$$

π reprezintă proporția sau ponderea unităților din populație care îndeplinesc o anumită proprietate sau condiție A, unde $0 \leq \pi \leq 1$.

Rezultă că:

$$\pi = \frac{M}{N}$$

unde M este numărul unităților din populație care îndeplinesc condiția A. Parametrul se poate exprima ca o valoare relativă sau în procente.

Se numește proporție de selecție statistica definită prin relația:

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

unde variabilele aleatoare de selecție X_i sunt variabile aleatoare de tip Bernoulli, identic repartizate și independente probabilistic, adică sunt definite prin tabloul:

$$X_i: \begin{pmatrix} 0 & 1 \\ 1-\pi & \pi \end{pmatrix}$$

Rezultă că:

$$E(X_i) = E(X) = \pi$$

$$V(X_i) = V(X) = \pi(1-\pi)$$

La nivelul unui eșantion extras după metoda sondajului aleator simplu, se definește proporția de eșantion:

$$w = \frac{m}{n}$$

unde m reprezintă numărul unităților din eșantion care îndeplinesc proprietatea A. Această valoare reprezintă estimația punctuală a parametrului π.

Proprietăți

Proporția de selecție este un estimator nedeplasat

Ca și în cazul mediei, se poate demonstra ușor că $E(\hat{\pi}) = \pi$, adică media estimatorului este egal cu parametrul π.

Proporția de selecție este un estimator convergent

Pentru un eșantion cu volum suficient de mare are loc:

$$(\hat{\pi}_n)_{n\in N} \xrightarrow{p} \pi$$

Primele două proprietăți asigură baza pentru estimarea punctuală a parametrului π.

Proporția de selecție este un estimator eficient

Notăm varianța mediei de selecție:

$$V(\hat{\pi}) = \sigma_{\hat{\pi}}^2$$

Aceasta se calculează diferențiat în funcție de tipul de extragere utilizat și asigură condiția de eficiență:

- în cazul extragerii aleatoare cu revenire

$$\sigma_{\hat{\pi}}^2 = \frac{V(X)}{n} = \frac{\pi(1-\pi)}{n}$$

- în cazul extragerii aleatoare fără revenire

$$\sigma_{\hat{\pi}}^2 = \frac{\pi(1-\pi)}{n}(1-f)$$

unde

$$(1-f) = \frac{N-n}{N}$$

Convergența în repartiție

Dacă standardizăm proporția de selecție, iar volumul eșantionului crește peste o anumită limită, se obține o variabilă aleatoare care urmează o lege de distribuție normală standard:

$$\left(\frac{\hat{\pi}-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right)_{n\in N} \xrightarrow{rep} Z \sim N(0,1)$$

2.2.5. Varianța de selecție

Definiție

Se numește varianță de selecție statistica:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

unde variabilele aleatoare de selecție X_i sunt variabile aleatoare identic repartizate și independente probabilistic.

O altă relație pentru varianța de selecție este:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\hat{\mu} + \hat{\mu}^2\right) = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\hat{\mu}\frac{1}{n}\sum_{i=1}^{n}X_i + \hat{\mu}^2$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \hat{\mu}^2 = \hat{\mu}_2 - \hat{\mu}^2$$

O valoare posibilă a acestui estimator este varianța de eșantion:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Proprietăți

Varianța de selecție este un estimator deplasat pentru parametrul σ^2

Se poate demonstra ușor că:

$$E(\hat{\sigma}^2) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2\right) = \frac{n-1}{n}\sigma^2$$

sau

$$E(\hat{\sigma}^2) = \sigma^2 - \frac{\sigma^2}{n}$$

Rezultă că deplasarea estimatorului este

$$B(\sigma) = -\frac{\sigma^2}{n}$$

Varianța estimatorului este:

$$V(\hat{\sigma}^2) = \frac{\sigma^{(4)} - \sigma^4}{n} - 2\frac{\sigma^{(4)} - \sigma^4}{n^2} + \frac{\sigma^{(4)} - 3\sigma^4}{n^3}$$

unde

$$\sigma^{(4)} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4$$

este momentul centrat de ordinul 4.

Varianța de selecție modificată

Pentru a corecta proprietatea de deplasare a variației de selecție, se construiește un alt estimator numit varianță de selecție modificată:

$$\hat{\sigma}'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

sau

$$\hat{\sigma}'^2 = \frac{n}{n-1}\hat{\sigma}^2$$

Se observă imediat că acest estimator este nedeplasat pentru parametrul varianță σ^2:

$$E(\hat{\sigma}'^2) = E\left(\frac{n}{n-1}\hat{\sigma}^2\right) = \frac{n}{n-1}E(\hat{\sigma}^2) = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2$$

Pentru un eșantion fixat, o valoare a statisticii este varianța de eșantion modificată și reprezintă estimația punctuală a varianței σ^2:

$$s'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

2.2.6. Moment centrat de selecție de ordin *k*

Pentru parametrul moment centrat de ordinul k se definește estimatorul:

$$\hat{\sigma}^{(k)} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^k$$

La nivelul unui eșantion fixat, se obține estimația punctuală:

$$s^{(k)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

2.3. Selecții pentru populații care urmează legi de distribuție normale

Considerăm o variabilă aleatoare care urmează o lege de distribuție normală de forma:

$$X \sim N(\mu, \sigma^2)$$

O selecție de volum n din populația de referință este dată prin vectorul aleator V: $(X_1, X_2, \ldots, X_n)$, unde $X_i \sim N(\mu, \sigma^2)$.

Pentru aceste selecții prezentăm o serie de rezultate mai importante care sunt utile în rezolvarea problemelor de inferență statistică.

Suma variabilelor aleatoare de selecție la pătrat reprezintă o variabilă aleatoare care urmează o lege de repartiție Chi-pătrat

Fără a restrânge generalitatea, considerăm $X_i \sim N(0, \sigma^2)$. Atunci,

$$Y = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n, \sigma)$$

Distribuția Chi-pătrat este dată prin funcția densitate:

$$f(x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\sigma^n \Gamma\left(\frac{n}{2}\right)} \cdot x^{\frac{n}{2}-1} e^{-\frac{x}{2\sigma^2}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

iar

$$\Gamma(a) = \int_0^{\infty} x^{a-1} e^{-x} dx,\ a > 0$$

Parametrii distribuției sunt n, volumul eșantionului (numărul de variabile sumate), care poartă numele de număr de grade de libertate, și abaterea standard σ.

Analog, dacă se consideră $X_i \sim N(\mu, \sigma^2)$,

$$Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n, 1)$$

În cazul în care se utilizează estimatorul variației de selecție modificate $\hat{\sigma}'^2$, are loc relația:

$$Y = \frac{(n-1)}{\sigma^2} \hat{\sigma}'^2 \sim \chi^2(n-1,1)$$

Suma unor variabile aleatoare Chi-pătrat reprezintă tot o variabilă care urmează legea Chi-pătrat

Fie $X_i \sim \chi^2(k_i, \sigma)$, atunci

$$Y = \sum_{i=1}^{n} X_i \sim \chi^2\left(\sum_{i=1}^{n} k_i, \sigma\right)$$

Raportul dintre o variabilă aleatoare normală și rădăcina pătrată a unei variabile Chi-pătrat reprezintă o variabilă Student

Fie $X \sim N(0, \sigma^2)$, $Y \sim \chi^2(n, \sigma)$, atunci

$$t = \frac{X}{\sqrt{\frac{Y}{n}}} \sim t(n)$$

Funcția densitate a repartiției Student este:

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi} \cdot \Gamma\left(\frac{n}{2}\right)} \cdot \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \qquad x \in R$$

Parametrul acestei distribuții este n, care poartă numele de număr de grade de libertate.

Dacă se consideră cazul estimatorului $\hat{\sigma}'^2$, cu proprietatea:

$$\frac{(n-1)}{\sigma^2}\hat{\sigma}'^2 \sim \chi^2(n-1,1)$$

se obține variabila Student utilizată în procesul de estimare:

$$t = \frac{\hat{\mu}-\mu}{\sqrt{\frac{\hat{\sigma}'^2}{n}}} \sim t(n-1)$$

Raportul a două variabile aleatoare Chi-pătrat reprezintă o variabilă aleatoare Fisher

Fie două variabile $U \sim \chi^2(m,\sigma)$, $V \sim \chi^2(n,\sigma)$, atunci:

$$X = \frac{\frac{U}{m}}{\frac{V}{n}} \sim F(m,n)$$

Pentru legea de repartiție Fisher, funcția densitate este:

$$f(x) = \begin{cases} \frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)} \cdot x^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}, & x>0 \\ 0, & x \le 0 \end{cases}$$

unde

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx, \qquad a>0, b>0$$

Cei doi parametri ai repartiției sunt cele două grade de libertate: n și m.

Dacă se iau în calcul două variabile aleatoare $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, iar pentru estimatorii varianțelor se consideră:

$$\frac{(n_1-1)}{\sigma_1^2}\hat{\sigma}_1'^2 \sim \chi^2(n_1-1,1)$$

$$\frac{(n_2-1)}{\sigma_2^2}\hat{\sigma}_2'^2 \sim \chi^2(n_2-1,1)$$

atunci

$$Y = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{\hat{\sigma}_1'^2}{\hat{\sigma}_2'^2} \sim F(n_1 - 1, n_2 - 1)$$

Suma sau diferența unui număr de variabile aleatoare normale reprezintă o variabilă aleatoare normală

Considerăm $X_i \sim N(\mu_i, \sigma_i^2)$ și n constante reale a_i. Are loc relația:

$$X = \sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

În particular, dacă avem două variabile aleatoare $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, atunci diferența lor urmează o lege normală de forma:

$$X = X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

În cazul variabilelor aleatoare de selecție $X_i \sim N(\mu, \sigma^2)$, are loc:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Evident,

$$Z = \frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Pentru două eșantioane de volume n_1 și n_2 și două variabile $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, se poate considera diferența dintre estimatorii mediilor, știind că:

$$\hat{\mu}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \ \hat{\mu}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Are loc:

$$\hat{\mu}_1 - \hat{\mu}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

iar

$$Z = \frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

În situația în care variațele celor două distribuții sunt egale, se obține variabila:

$$Z = \frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

Dacă nu se cunosc variațele celor două variabile, se construiește variabila Student:

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{\sigma}_1'^2}{n_1} + \frac{\hat{\sigma}_2'^2}{n_2}}} \sim t(n_1 + n_2 - 2)$$

În cazul în care cele două distribuții au o varianță egală și necunoscută, $X_1 \sim N(\mu_1, \sigma^2)$, $X_2 \sim N(\mu_2, \sigma^2)$, pentru diferența dintre estimatorii mediei se construiește variabila Student:

$$t = \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \cdot \frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sqrt{(n_1 - 1)\hat{\sigma}_1'^2 + (n_2 - 1)\hat{\sigma}_2'^2}} \sim t(n_1 + n_2 - 2)$$

2.4. Estimarea prin interval de încredere a unui parametru

Estimarea prin interval de încredere are la bază ideea de a construi un interval aleator care să conțină, cu o anumită probabilitate cunoscută numită *nivel de încredere*, parametrul estimat.

Considerăm $\alpha \in (0,1)$, cu α aproape de zero. Notăm $(1 - \alpha)$ nivelul de încredere sau probabilitatea cu care estimăm parametrul.

A estima prin interval de încredere un parametru θ înseamnă a determina două statistici, numite *limita inferioară* (L_i) și *limita superioară* (L_s), astfel încât să aibă loc relația:

$$P(L_i \leq \theta \leq L_s) = (1 - \alpha)$$

Observația 1

Deoarece parametrul este o constantă și limitele intervalului sunt variabile, este corect ca la interpretare să se spună că intervalul *acoperă* parametrul (sau că intervalul conține parametrul) cu probabilitatea $(1 - \alpha)$. Evident, α este probabilitatea ca intervalul să nu acopere sau să nu conțină parametrul.

Observația 2

În cazul legilor de repartiție simetrice și unimodale, intervalele de încredere au o lungime minimă. Această proprietate se aplică în special în cazul estimării mediei și a proporției.

2.4.1. Estimarea mediei

Determinarea intervalului de încredere pentru parametrul medie necesită cunoașterea legii de repartiție a estimatorului. Există două situații:

- dacă se cunoaște legea de repartiție a variabilei analizate, de exemplu $X \sim N(\mu, \sigma^2)$, media de selecție are tot o lege probabilistă cunoscută, mai precis este vorba despre distribuția normală:

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- dacă nu se cunoaște legea de repartiție a variabilei X, se poate utiliza proprietatea de convergență pe care o oferă *teorema limită centrală*, adică:

$$\left(\frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)_{n \in N} \xrightarrow{rep} Z \sim N(0,1)$$

Așa cum se observă din relațiile de mai sus, legile de repartiție ale statisticilor depind nu numai de parametrul μ, ci și de σ^2. Rezultă două situații pentru estimarea prin interval de încredere a mediei.

a. Intervalul de încredere al mediei pentru cazul σ^2 cunoscut

În funcție de tipul de extragere a eșantionului (cu sau fără revenire), avem iarăși două situații. Practic, intervalul se obține la fel în ambele cazuri, ceea ce diferă fiind doar varianța mediei de selecție.

Intervalul de încredere în cazul extragerii aleatoare cu revenire

În acest caz, se utilizează statistica

$$Z = \frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

pentru care se poate accepta că urmează o lege de repartiție normală standard, cel puțin asimptotic.

Dacă $Z \sim N(0,1)$, se utilizează funcția densitate:

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}, \qquad x \in R$$

care are următoarea formă:

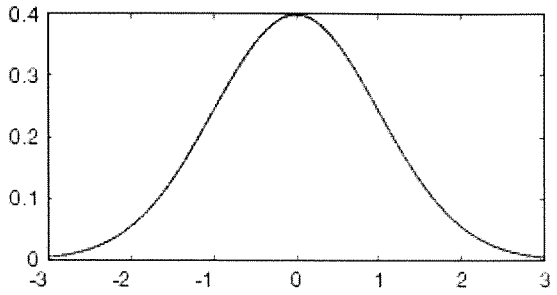


Figura 2.2. *Repartiția normală standard N(0,1)*

De asemenea, pentru stabilirea intervalului de încredere, se utilizează funcția lui Laplace:

$$\Phi(z) = \int_0^z \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} dt$$

Pentru un nivel de încredere egal cu $(1-\alpha)$, având o lege de distribuție simetrică, intervalul de încredere este minim și se obține din relația:

$$P(-z \leq Z \leq z) = (1-\alpha)$$

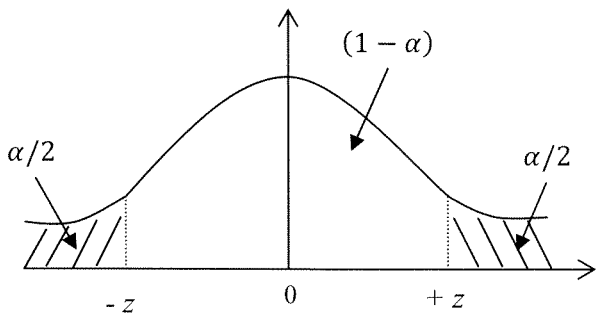


Figura 2.3. *Valorile teoretice ale distribuției normale standard pentru un nivel de încredere de $(1-\alpha)$*

Conform legii Gauss-Laplace, probabilitatea de mai sus se poate calcula cu ajutorul funcției lui Laplace (vezi figura 2.3):

$$P(-z \leq Z \leq z) = \int_{-z}^{z} f(t)\, dt = 2\Phi(z) = 1-\alpha$$

Folosind tabela lui Laplace, se poate obține z pe baza relației:

$$\Phi(z) = \frac{1-\alpha}{2}$$

Pentru obținerea intervalului de încredere, se fac câteva modificări elementare în relația de mai jos:

$$P(-z \leq Z \leq z) = P\left(-z \leq \frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z\right) = 1 - \alpha$$

Rezultă:

$$P\left(\hat{\mu} - z\frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu} + z\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Cele două statistici care determină limitele intervalului de încredere sunt:

$$L_i = \hat{\mu} - z\frac{\sigma}{\sqrt{n}}$$

$$L_s = \hat{\mu} + z\frac{\sigma}{\sqrt{n}}$$

iar intervalul poate fi scris astfel:

$$\left(\hat{\mu} \pm z\frac{\sigma}{\sqrt{n}}\right)$$

La nivelul unui eșantion extras aleator cu revenire din populație, se obține un interval de încredere estimat sau calculat de forma:

$$\left(\bar{x} \pm z\frac{\sigma}{\sqrt{n}}\right)$$

Observații

- lungimea intervalului de încredere este:

$$2 \cdot z\frac{\sigma}{\sqrt{n}}$$

- eroare maxim admisibilă sau eroare limită pentru parametrul medie reprezintă mărimea intervalului de încredere:

$$\Delta_\mu = z\frac{\sigma}{\sqrt{n}}$$

- eroarea limită cuprinde două componente care vizează nivelul de încredere și calitatea estimatorului;

- dacă se modifică probabilitatea $(1-\alpha)$, se modifică valoarea teoretică z, deci se modifică lungimea intervalului. De regulă, creșterea probabilității conduce la o creștere a mărimii intervalului și, prin urmare, o scădere a preciziei estimării;
- abaterea standard a estimatorului se mai numește și eroare standard a mediei (Standard Error - SE) și este dată de relația:

$$SE = \frac{\sigma}{\sqrt{n}}$$

- eroarea standard admite alte două componente care pot modifica mărimea intervalului de încredere: volumul eșantionului și gradul de omogenitate a populației, măsurat cu abaterea standard σ. Dacă volumul eșantionului crește, mărimea intervalului scade, adică se înregistrează o creștere a preciziei estimării. Evident, eșantionul nu poate fi mărit decât până la o anumită limită, în general, din rațiuni de timp și de cost. Întrebarea este dacă se poate micșora valoarea lui σ pentru a crește gradul de precizie. Pentru moment, răspunsul este da (de exemplu, printr-o grupare a unităților populației pe grupe omogene), iar explicațiile vor fi prezentate în capitolul dedicat sondajelor mixte;
- intervalul de încredere se poate scrie astfel:

$$(\bar{x} \pm z \cdot SE)$$

Exemplu

Într-o cercetare prin sondaj aleator simplu, extragere cu revenire, pentru un eșantion de $n = 400$ de studenți, s-a realizat un studiu privind utilizarea telefonului mobil. Pentru variabila (X) timp de convorbire pe zi exprimat în minute, la nivelul eșantionului, s-au obținut rezultatele: timp mediu de convorbire pe student de 125 de minute și o abatere standard de eșantion modificată de 25 de minute. Dintr-o cercetare anterioară, se cunoaște că abaterea standard la nivelul întregii populații este de 50 de minute. Să se estimeze prin interval de încredere numărul mediu de minute consumate de un student la nivelul întregii populații, pentru o încredere de 95% și pentru situația de 90%.

Intervalul de încredere se calculează cu ajutorul relației:

$$\left(\bar{x} \pm z\frac{\sigma}{\sqrt{n}}\right)$$

sau

$$(\bar{x} \pm z \cdot SE)$$

În relația de mai sus se cunosc: $\bar{x} = 125$; $n = 400$; $\sigma = 50$.

Valoarea lui z se citește din tabelul probabilistic al distribuției Laplace, pentru nivelul de încredere stabilit, conform relației:

$$\Phi(z) = \frac{1-\alpha}{2}$$

În cazul unui nivel de încredere de 0,95, se citește valoarea lui z corespunzătoare funcției lui Laplace:

$$\Phi(z) = \frac{0{,}95}{2} = 0{,}475$$

Din tabel rezultă: $z = 1{,}96$.

Intervalul de încredere pentru o probabilitate de 0,95 este:

$\left(125 \pm 1{,}96\frac{50}{\sqrt{400}}\right)$, adică $(125 \pm 1{,}96 \cdot 2{,}5)$ sau $(120{,}1;\ 129{,}9)$ minute.

Mărimea intervalului de încredere este de 4,9 minute, iar eroarea standard a mediei este $SE = \frac{50}{\sqrt{400}} = 2{,}5$ minute.

Interpretare

Numărul mediu de minute la telefon utilizate de un student din populația de referință este acoperit de intervalul $(120{,}1;\ 129{,}9)$ minute, cu o probabilitate de 0,95.

În cel de-al doilea caz, probabilitatea este mai mică și se utilizează relația:

$\Phi(z) = \frac{0{,}90}{2} = 0{,}45$, de unde rezultă valoarea $z = 1{,}65$.

După calcule, rezultă intervalul de încredere:

$\left(125 \pm 1{,}65\frac{50}{\sqrt{400}}\right)$, adică $(125 \pm 4{,}125)$, deci un interval mai mic de:

$$(120{,}875;\ 129{,}125) \text{ minute.}$$

Intervalul de încredere în cazul extragerii aleatoare fără revenire

Pentru acest tip de extragere se ține cont că varianța mediei de selecție este:

$$\sigma_{\hat{\mu}}^2 = \frac{\sigma^2}{n}(1-f)$$

iar intervalul se construiește pe baza relației:

$$P\left(\hat{\mu} - z\sqrt{\frac{\sigma^2}{n}(1-f)} \le \mu \le \hat{\mu} + z\sqrt{\frac{\sigma^2}{n}(1-f)}\right) = 1-\alpha$$

adică este de forma:

$$\left(\hat{\mu} \pm z\sqrt{\frac{\sigma^2}{n}(1-f)}\right)$$

La nivelul unui eșantion posibil de extras, se obține relația de calcul pentru intervalul estimat:

$$\left(\bar{x} \pm z\sqrt{\frac{\sigma^2}{n}(1-f)}\right)$$

Observație

Deoarece abaterea standard a estimatorului mediei este

$$\sigma \cdot \sqrt{\frac{(1-f)}{n}}$$

atunci coeficientul de variație al mediei de selecție are relația:

$$\frac{\sigma}{\mu} \cdot \sqrt{\frac{(1-f)}{n}}$$

Însă raportul $\sigma/\mu = v$ este tocmai coeficientul de variație la nivelul populației totale. Prin urmare, coeficientul de variație al mediei de selecție este:

$$v_{\hat{\mu}} = v \cdot \sqrt{\frac{(1-f)}{n}}$$

Exemplu

Considerăm datele de la exemplul anterior, o probabilitate de 0,95 și, în plus, se consideră o extragere fără revenire dintr-o populație de 10000 de studenți. Fracția de sondaj este:

$$f = \frac{n}{N} = \frac{400}{10000} = 0{,}04$$

În acest caz, intervalul de încredere se calculează astfel:

$$\left(\bar{x} \pm z\sqrt{\frac{\sigma^2}{n}(1-f)}\right) = \left(125 \pm 1{,}96\sqrt{\frac{50^2}{400}(1-0{,}04)}\right) = (125 \pm 4{,}8)$$

minute.

Observație

Față de cazul extragerii cu revenire, intervalul de încredere este mai mic, pentru aceeași probabilitate, adică rezultatul este mai precis. Diferența dintre cele două rezultate devine nesemnificativă dacă volumul populației este unul foarte mare.

b. Intervalul de încredere al mediei pentru cazul σ^2 necunoscut

Intervalul de încredere în cazul extragerii aleatoare cu revenire

În cele mai multe cazuri, parametrul varianță nu este cunoscut și trebuie estimat. Dacă se utilizează ca estimator varianța de selecție modificată, pentru un eșantion suficient de mare, se obține statistica Student:

$$t = \frac{\hat{\mu} - \mu}{\sqrt{\frac{\hat{\sigma}'^2}{n}}} \sim t(n-1)$$

Repartiția Student este simetrică, deci se poate construi un interval de încredere minim pentru un nivel de încredere cunoscut $(1-\alpha)$. Cu alte cuvinte, folosind tabela Student, se poate determina o valoare $t_{\alpha/2}$ care să respecte condiția:

$$P\left(-t_{\alpha/2} \leq t \leq t_{\alpha/2}\right) = P\left(-t_{\alpha/2} \leq \frac{\hat{\mu} - \mu}{\frac{\hat{\sigma}'}{\sqrt{n}}} \leq t_{\alpha/2}\right) = 1 - \alpha$$

Tabelele probabiliste pentru repartiția Student oferă valoarea teoretică în baza condiției:

$$P\left(t > t_{\alpha/2}\right) = \frac{\alpha}{2}$$

Rezultă imediat:

$$P\left(\hat{\mu} - t_{\alpha/2}\frac{\hat{\sigma}'}{\sqrt{n}} \leq \mu \leq \hat{\mu} + t_{\alpha/2}\frac{\hat{\sigma}'}{\sqrt{n}}\right) = 1 - \alpha$$

adică intervalul de încredere este de forma:

$$\left(\hat{\mu} \pm t_{\alpha/2}\frac{\hat{\sigma}'}{\sqrt{n}}\right)$$

La nivelul unui eșantion extras din populație, se obține relația de calcul pentru intervalul estimat:

$$\left(\bar{x} \pm t_{\alpha/2}\frac{s'}{\sqrt{n}}\right)$$

Intervalul de încredere în cazul extragerii aleatoare fără revenire

Se utilizează relația:

$$P\left(\hat{\mu} - t_{\alpha/2}\sqrt{\frac{\hat{\sigma}'^2}{n}(1-f)} \leq \mu \leq \hat{\mu} + t_{\alpha/2}\sqrt{\frac{\hat{\sigma}'^2}{n}(1-f)}\right) = 1 - \alpha$$

de unde rezultă intervalul:

$$\left(\hat{\mu} \pm t_{\alpha/2}\sqrt{\frac{\hat{\sigma}'^2}{n}(1-f)}\right)$$

Pentru un eșantion fixat de volum n se obține intervalul calculat:

$$\left(\bar{x} \pm t_{\alpha/2}\sqrt{\frac{s'^2}{n}(1-f)}\right)$$

Exemplu

Considerăm datele de la exemplul precedent, fără a dispune însă de o valoare a abaterii standard la nivelul populației totale, și se consideră un nivel de încredere de 0,95. Se cunoaște însă $s'^2 = 25$ de minute. În aceste condiții, intervalul de încredere se determină pe baza abaterii standard de eșantion modificate.

În cazul extragerii cu revenire, intervalul de încredere pentru medie se obține cu relația:

$$\left(\bar{x} \pm t_{\alpha/2}\frac{s'}{\sqrt{n}}\right)$$

Valoarea teoretică se citește din tabelul Student pentru relația:
$P(t > t_{\alpha/2}) = 0{,}025$. Rezultă o valoare $t_{\alpha/2} = 1{,}96$, iar intervalul este:
$\left(125 \pm 1{,}96\frac{50}{\sqrt{400}}\right)$, adică $(125 \pm 4{,}9)$ minute.

2.4.2. Estimarea valorii globale (a totalului)

Pentru parametrul $\tau = \sum_{i=1}^{N} x_i$ sau $\tau = N \cdot \mu$ se utilizează estimatorul nedeplasat și convergent:

$$\hat{\tau} = N \cdot \hat{\mu} = \frac{N}{n}\sum_{i=1}^{n} X_i$$

La nivel de eșantion, se estimează punctual totalul prin estimația:

$$T = N \cdot \bar{x} = \frac{N}{n}\sum_{i=1}^{n} x_i$$

Ca și în cazul mediei, există cele două situații, când se cunoaște sau nu parametrul varianței. Exemplificăm doar cazul cel mai întâlnit, când varianța este necunoscută.

Intervalul de încredere în cazul extragerii aleatoare cu revenire

$$\left(\hat{\tau} \pm t_{\alpha/2}\hat{\sigma}_{\hat{\tau}}\right)$$

sau

$$\left(\hat{\tau} \pm t_{\alpha/2}\frac{N}{\sqrt{n}}\hat{\sigma}'\right)$$

La nivelul unui eșantion extras, se calculează intervalul:

$$\left(T \pm t_{\alpha/2}\frac{N}{\sqrt{n}}s'\right)$$

Intervalul de încredere în cazul extragerii aleatoare fără revenire

Se construiește intervalul de încredere:

$$\left(\hat{\tau} \pm t_{\alpha/2}\sqrt{N(N-n)\frac{\hat{\sigma}'^2}{n}}\right)$$

La nivelul unui eșantion extras, se calculează intervalul:

$$\left(T \pm t_{\alpha/2}\sqrt{N(N-n)\frac{s'^2}{n}}\right)$$

Exemplu

Pentru datele de la exemplul precedent, cunoscând că volumul populației studenților este de 10000, să se estimeze prin interval de încredere numărul total de minute consumate la nivelul întregii populații, cu o probabilitate de 0,95.

La nivel de eșantion, intervalul de încredere este:

$$\left(T \pm t_{\alpha/2}\sqrt{N(N-n)\frac{s'^2}{n}}\right)$$

Valoarea totală la nivel de eșantion este $T = N \cdot \bar{x} = 10000 \cdot 125 = 1250000$.

Din tabelul Student, valoarea teoretică, pentru o încredere de 95%, este 1,96.

Rezultă intervalul:

$\left(1250000 \pm 1{,}96\sqrt{10000 \cdot 9600 \cdot \frac{25}{400}}\right)$, adică (1250000 ± 4800) minute.

2.4.3. Estimarea proporției

Pentru parametrul proporție, procedeul de estimare prin interval de încredere urmează pașii prezentați pentru cazul mediei. Pentru fiecare caz în parte, prezentăm în sinteză relațiile estimatorului proporție de selecție:

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

a. Intervalul de încredere al proporției pentru cazul σ^2 cunoscut

Cazul sondajului aleator cu revenire

Se utilizează statistica

$$Z = \frac{\hat{\pi} - \pi}{\frac{\sigma}{\sqrt{n}}}$$

pentru care se acceptă ipoteza că tinde spre o repartiție normală standard.

Intervalul de încredere, pentru o probabilitate $(1 - \alpha)$, este:

$$\left(\hat{\pi} \pm z\frac{\sigma}{\sqrt{n}}\right)$$

unde z se citește din tabela Laplace, pentru

$$\Phi(z) = \frac{1 - \alpha}{2}$$

Interval estimat corespunzător este:

$$\left(w \pm z\frac{\sigma}{\sqrt{n}}\right)$$

sau

$$(w \pm z \cdot SE)$$

Eroarea standard a proporției se mai poate scrie:

$$SE = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Observație

Mărimea $\Delta_\pi = z\frac{\sigma}{\sqrt{n}}$ se numește eroare limită de estimare a proporției. Această limită se poate fixa a priori pentru a putea apoi calcula un volum al eșantionului care să asigure o estimație a parametrului în limitele acestei erori fixate.

Cazul sondajului aleator fără revenire

Intervalul prezentat în cazul precedent se modifică la nivelul variației estimatorului cu coeficientul de corecție $(1-f)$:

$$\left(\hat{\pi} \pm z\sqrt{\frac{\sigma^2}{n}(1-f)}\right)$$

Similar, intervalul estimat la nivelul unui eșantion este:

$$\left(w \pm z\sqrt{\frac{\sigma^2}{n}(1-f)}\right)$$

Exemplu

Considerăm cazul unui eșantion de 100 de studenți extrași aleator simplu fără revenire dintr-o populație de 5000 de studenți. Din anchetă se observă că, la nivel de eșantion, numărul studenților care dețin un smartphone este egal cu 85. Se cere să se estimeze prin interval de încredere procentul studenților care dețin un smartphone la nivelul întregii populații, cu o încredere de 98%, considerând că dintr-un studiu anterior se cunoaște o varianță a variabilei analizate în valoare de 0,16.

Pentru calculul intervalului de încredere se determină mai întâi ponderea estimată de studenți care au un smartphone:

$$w = \frac{85}{100} = 0{,}85$$

Din tabela Laplace, se citește o valoare pentru

$$\Phi(z) = \frac{1-\alpha}{2} = 0{,}49$$

Rezultă valoarea $z = 2{,}33$.

Fracția de sondaj este:

$$f = \frac{n}{N} = \frac{100}{5000} = 0{,}02$$

Intervalul de încredere este:

$$\left(w \pm z\sqrt{\frac{\sigma^2}{n}(1-f)}\right) = (0{,}85 \pm 2{,}33 \cdot 0{,}039) = (0{,}85 \pm 0{,}09)$$

În procente, intervalul este: $(85\% \pm 9\%)$ sau $(76\%; 94\%)$.

Ca interpretare, se poate aprecia că procentul real al studenților care dețin un smartphone este acoperit de intervalul $(76\%; 94\%)$, cu o probabilitate de 0,98.

b. Intervalul de încredere al proporției pentru cazul σ^2 necunoscut

Cazul sondajului aleator cu revenire

Determinarea intervalului de încredere presupune mai întâi identificarea estimatorul varianței:

$$\hat{\sigma}_\pi^2 = \frac{\hat{\pi}(1-\hat{\pi})}{n}$$

Statistica utilizată în acest caz este:

$$t = \frac{\hat{\pi} - \pi}{\frac{\hat{\sigma}_\pi}{\sqrt{n}}}$$

pentru care se acceptă că tinde spre o repartiție Student de $n - 1$ grade de libertate.

Intervalul de încredere, pentru o probabilitate $(1 - \alpha)$, este:

$$\left(\hat{\pi} \pm t_{\alpha/2}\frac{\hat{\sigma}_\pi}{\sqrt{n}}\right)$$

sau

$$\left(\hat{\pi} \pm t_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)$$

Valoarea teoretică din tabelul Student se obține pentru:

$$P(t > t_{\alpha/2}) = \frac{\alpha}{2}$$

Interval estimat se obține pe baza estimației punctuale a proporție w:

$$\left(w \pm t_{\alpha/2} \frac{s_\pi}{\sqrt{n}}\right)$$

sau

$$\left(w \pm t_{\alpha/2} \sqrt{\frac{w(1-w)}{n}}\right)$$

Eroarea standard estimată a parametrului proporție este de forma:

$$SE = \sqrt{\frac{w(1-w)}{n}}$$

Cazul sondajului aleator fără revenire

Intervalul probabilist are relația:

$$\left(\hat{\pi} \pm t_{\alpha/2} \sqrt{\frac{\hat{\sigma}_\pi^2}{n}(1-f)}\right)$$

iar intervalul la nivel de eșantion este:

$$\left(w \pm t_{\alpha/2} \sqrt{\frac{w(1-w)}{n}(1-f)}\right)$$

Exemplu

Pentru datele din exemplul anterior, se consideră doar rezultatele de la nivelul eșantionului pentru estimarea proporției prin interval de încredere.

În acest caz, se utilizează relația:

$$\left(w \pm t_{\alpha/2} \sqrt{\frac{w(1-w)}{n}(1-f)}\right)$$

Din calcule, rezultă intervalul:

$(0{,}85 \pm 2{,}33 \cdot 0{,}034)$ sau $(0{,}85 \pm 0{,}079)$.

2.4.3. Estimarea variației

Analizăm această problemă doar în cazul sondajului aleator simplu cu revenire. Pentru o variabilă care urmează legea normală, repartiția admite doi parametri: media și varianța. Determinarea intervalului de încredere al variației presupune două situații: se cunoaște sau nu parametrul medie.

a. Intervalul de încredere al varianței pentru cazul μ cunoscut

În condițiile în care considerăm că variabila X urmează o lege normală, se utilizează o statistică de forma:

$$H = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n, 1)$$

Pentru un nivel de încredere $(1 - \alpha)$ fixat se utilizează repartiția Chi-pătrat pentru a determina două mărimi h_1 și h_2 pentru care are loc:

$$P(h_1 \leq H \leq h_2) = (1 - \alpha)$$

Aceste valori se citesc din tabela Chi-pătrat pe baza condițiilor:

$$P(H > h_1) = 1 - \frac{\alpha}{2}$$

$$P(H > h_2) = \frac{\alpha}{2}$$

Intervalul de încredere se obține pe baza relației:

$$P\left(h_1 \leq \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \leq h_2\right) = (1 - \alpha)$$

Rezultă:

$$P\left(\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{h_2} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{h_1}\right) = (1 - \alpha)$$

Intervalul de încredere calculat la nivelul unui eșantion este de forma:

$$\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{h_2}; \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{h_1}\right)$$

b. Intervalul de încredere al varianței pentru cazul μ necunoscut

Dacă parametrul medie nu este cunoscut, se utilizează o altă statistică:

$$H = \frac{(n-1)}{\sigma^2}\hat{\sigma}'^2 \sim \chi^2(n - 1{,}1)$$

Similar cu cazul anterior, intervalul de încredere se bazează pe proprietatea:

$$P(h_1 \leq H \leq h_2) = (1 - \alpha)$$

iar valorile teoretice se obțin pe baza tabelelor probabiliste pentru:

$$P(H > h_1) = 1 - \frac{\alpha}{2}$$

$$P(H > h_2) = \frac{\alpha}{2}$$

Intervalul de încredere probabilist este următorul:

$$\left(\frac{(n-1)\hat{\sigma}'^2}{h_2}; \frac{(n-1)\hat{\sigma}'^2}{h_1}\right)$$

iar cel estimat la nivel de eșantion este:

$$\left(\frac{(n-1)s'^2}{h_2}; \frac{(n-1)s'^2}{h_1}\right)$$

2.5. Calculul volumului eșantionului

Sondajele aleatoare simple impun o condiție foarte importantă, anume cunoașterea volumului eșantionului. Cele mai multe cercetări prin sondaj încep cu această întrebare: cât de mare trebuie să fie eșantionul? Am precizat deja că eșantioanele aleatoare impun o condiție de volum, adică sunt eșantioane mari. De asemenea, am arătat că precizia estimațiilor obținute poate fi ridicată dacă se crește volumul eșantionului.

În funcție de situațiile existente în practică, putem sublinia două posibilități privind calculul volumului eșantionului: dacă există condiții de costuri impuse și dacă există limite privind erorile cu care sunt obținute rezultatele.

Așadar, prima situație se referă la existența unor restricții de cost impuse a priori. În acest caz, se împarte suma totală disponibilă (costul total - c_T) la costul unitar (c_u) pe chestionar și se obține mărimea eșantionului:

$$n = \frac{c_T}{c_u}$$

Pentru a calcula volumul eșantionului este necesar să se cunoască informația privitoare la costul unitar. De regulă, această informație provine din studii anterioare sau se poate estima prin apelul la rezultatele din anchete similare.

A doua situație este legată de utilizarea unei condiții privind precizia rezultatelor. Asemenea limite de precizie se pot stabili prin utilizarea mai multor măsuri cu privire la erorile de eșantionare: varianța estimatorului, mărimea intervalului de încredere, eroarea maxim admisibilă.

Cel mai frecvent utilizată este metoda care permite calculul unui anumit volum al eșantionului care să corespundă unei erori de estimare asumate și a unei

probabilități fixate a priori. Se stabilește o anumită limită sau o valoare a *erorii maxim admisibile* pentru un parametru care se dorește a fi estimat cu ajutorul sondajului și se fixează și probabilitatea cu care este estimat acel parametru.

Vom arăta în continuare cum se realizează acest calcul în funcție de variabila de interes utilizată și de parametrul luat în considerare. Ne referim la două situații: pentru o variabilă binară și parametrul proporție; pentru o variabilă numerică și parametrul medie.

2.5.1. Cazul unei variabile alternative

Considerăm o variabilă binară:

$$X: \begin{pmatrix} 0 & 1 \\ 1-\pi & \pi \end{pmatrix}$$

Obiectivul sondajului este acela de a estima parametrul π în condiții de calitate fixate a priori.

Deoarece măsurile cu privire la erori (varianța estimatorului, eroarea maxim admisibilă sau eroarea limită) depind de tipul de extragere, calculul se realizează distinct pentru sondajul aleator cu revenire și cel fără revenire.

a. Sondajul aleator cu revenire

În acest caz, eroarea maxim admisibilă pentru proporție este de forma:

$$\Delta_\pi = z\frac{\sigma}{\sqrt{n}}$$

Δ_π este fixat în funcție de natura variabilei (de exemplu, pentru anchetele de opinie, specialiștii fixează eroarea limită undeva între 1% și 2%).

Rămâne să se stabilească ipotezele pentru necunoscutele z și σ.

Valoarea lui z se citește din tabela Laplace pentru nivelul de încredere $(1-\alpha)$ fixat de cercetător. De exemplu, pentru o probabilitate de 0,95 se obține $z = 1{,}96$.

Parametrul σ este, de regulă, necunoscut. Pentru acest parametru se poate utiliza o valoare dintr-o cercetare anterioară (acceptând ipoteza că la nivelul populației studiate nu au avut loc schimbări structurale semnificative). O altă ipoteză frecvent utilizată în cercetare este aceea de a considera valoarea maximă pentru varianță, ceea ce corespunde situației de maximă eterogenitate a populației sau situației celei mai nefavorabile. Pentru o variabilă binară, varianța este maximă dacă se acceptă ipoteza că parametrul proporție ia valoarea 0,5.

Rezultă că $\sigma^2 = 0{,}25$.

Având aceste condiții, se poate calcula volumul eșantionului pe baza relației:

$$n = \frac{z^2\sigma^2}{\Delta_\pi^2}$$

Exemplu

Dacă se consideră o eroare maxim admisibilă de 2%, o probabilitate de 0,95 și o varianță maximă, atunci volumul eșantionului care estimează proporția populației în aceste condiții date este:

$$n = \frac{1{,}96^2 \cdot 0{,}25}{0{,}02^2} = 2401 \text{ unități.}$$

Dacă se mărește eroarea maxim admisibilă la 3%, ne așteptăm ca volumul eșantionului să scadă:

$$n = \frac{1{,}96^2 \cdot 0{,}25}{0{,}03^2} = 1067 \text{ de unități.}$$

În egală măsură, dacă se modifică varianța la nivelul populației, în sensul micșorării ei sub valoarea maximă (0,25), se micșorează volumul eșantionului. De exemplu, la o varianță de 0,16 și la o eroare maxim admisibilă de 2%, volumul eșantionului devine:

$$n = \frac{1{,}96^2 \cdot 0{,}16}{0{,}02^2} = 1537 \text{ de unități.}$$

Calculul volumului eșantionului se poate realiza și prin stabilirea unei limite cu privire la varianța estimatorului parametrului proporție. De exemplu, dacă nivelul variației nu trebuie să depășească o anumită limită F, atunci volumul eșantionului se poate obține astfel:

$$V(\hat{\pi}) = \frac{\sigma}{\sqrt{n}} \leq \mathcal{F}$$

Rezultă:

$$\frac{\sigma^2}{n} = \mathcal{F}$$

de unde

$$n = \frac{\sigma^2}{\mathcal{F}}$$

b. Sondajul aleator fără revenire

Față de cazul anterior, calculul se schimbă pentru că se ține cont de mărimea populației. Eroarea maxim admisibilă este dată prin relația:

$$\Delta_\pi = z\sqrt{\frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}}$$

Pentru simplificarea calculului, considerăm:

$$\Delta_\pi = z\sqrt{\frac{\sigma^2}{n}\cdot\frac{N-n}{N}} = z\sqrt{\frac{\sigma^2}{n}\cdot\left(1-\frac{n}{N}\right)}$$

Efectuând calculele necesare, se obține relația:

$$n = \frac{z^2\sigma^2}{\Delta_\pi^2 + \frac{z^2\sigma^2}{N}}$$

sau

$$n = \frac{z^2\sigma^2 N}{N\Delta_\pi^2 + z^2\sigma^2}$$

Dacă notăm cu

$$n_0 = \frac{z^2\sigma^2}{\Delta_\pi^2}$$

volumul eșantionului din cazul extragerii cu revenire se poate scrie:

$$n = \frac{z^2\sigma^2}{\Delta_\pi^2 + \frac{z^2\sigma^2}{N}} = \frac{\sigma^2}{\frac{\sigma^2}{N} + \frac{\Delta_\pi^2}{z^2}} = \frac{1}{\frac{1}{N} + \frac{\Delta_\pi^2}{z^2\sigma^2}} = \frac{1}{\frac{1}{N} + \frac{1}{n_0}}$$

sau

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Dacă $N \longrightarrow \infty$ este clar că $n = n_0$.

Se poate observa că, în cazul sondajului aleator fără revenire, volumul eșantionului este mai redus decât în cazul sondajului cu revenire. Relația de calul de mai sus este valabilă și dacă se consideră o condiție pentru varianța estimatorului, doar că se folosește relația de bază:

$$n_0 = \frac{\sigma^2}{\mathcal{F}}$$

Exemplu

Pentru o eroare limită de 2% și o populație de 10 milioane de unități se obține un eșantion: $n = 2400$ de unități. Dacă se ia în calcul exemplul anterior, se observă că, pentru un volum al populației foarte mare, diferența dintre cele două tipuri de extragere, cu sau fără revenire, devine nesemnificativă, adică volumul eșantioanelor obținute prin cele două metode este aproximativ același.

În schimb, dacă populația are un volum mai redus, situația se schimbă. Să considerăm, spre exemplu, că pentru o populație de 100 de mii de unități se obține un eșantion de 2345 de unități, evident, mai redus decât în cazul sondajului aleator cu revenire.

2.5.2. Cazul unei variabile numerice continue

Se consideră că variabila de interes este o variabilă numerică continuă care la nivelul populației urmează aproximativ o lege normală: $X \sim N(\mu, \sigma^2)$. Parametrul de interes este reprezentat de media populației, iar asupra varianței se adoptă o anumită ipoteză.

Pentru calculul volumului eșantionului care să estimeze parametrul medie în anumite condiții de eroare și de probabilitate fixate, se utilizează aceleași relații prezentate anterior pentru cazul proporției. Singura diferență în acest caz este legată de obținerea unei informații cu privire la omogenitatea populației sau o valoare pentru parametrul σ.

Opțiunea cea mai bună este să se folosească o informație din studii anterioare sau să se realizeze o investigație mai restrânsă care să permită evaluarea nivelului acestui parametru.

Așadar, pentru cazul unui eșantion extras aleator simplu cu revenire, volumul eșantionului se calculează astfel:

$$n = \frac{z^2\sigma^2}{\Delta_\mu^2}$$

unde $\sigma^2 = V(X)$, iar Δ_μ este eroarea maxim admisibilă pentru parametrul medie. Această eroare se stabilește cunoscând detaliile fenomenului analizat și reprezintă o fracțiune din abaterea standard.

Mult mai simplu pentru interpretare este să se ia în calcul o eroare standard pentru medie acceptabilă, în condiții de probabilitate asumate. În acest caz, se poate înlocui în relația de mai sus:

$$\Delta_\mu = z \cdot SE$$

Pentru o populație de volum redus, caz în care se aplică extragerea fără revenire, volumul eșantionului se calculează după relația:

$$n = \frac{z^2\sigma^2}{\Delta_{\bar{\mu}}^2 + \frac{z^2\sigma^2}{N}}$$

sau

$$n = \frac{1}{\frac{1}{N} + \frac{\Delta_{\bar{\mu}}^2}{z^2\sigma^2}} = \frac{1}{\frac{1}{N} + \frac{1}{n_0}} = \frac{n_0}{1 + \frac{n_0}{N}}$$

unde

$$n_0 = \frac{z^2\sigma^2}{\Delta_{\bar{\mu}}^2}$$

Exemplu

Considerăm cazul anchetei în care se studiază variabila (X) timp de convorbire pe zi, exprimat în minute, realizat de un student dintr-o populație de 10000 de persoane, cunoscând că varianța variabilei analizate este egală cu 2500. Se cere să se estimeze volumul eșantionului care trebuie extras aleator fără revenire, dacă se dorește să se estimeze timpul mediu de minute de convorbiri pe zi cu o eroare standard care nu depășește 1,2 minute pe student, cu o probabilitate de 0,95.

Utilizăm relațiile:

$$n = \frac{1}{\frac{1}{N} + \frac{1}{n_0}}$$

$$n_0 = \frac{z^2\sigma^2}{\Delta_{\bar{\mu}}^2} = \frac{\sigma^2}{SE^2}$$

Rezultă că $n_0 = 2083$ de studenți, iar $n = 1724$ de studenți.

2.6. Exerciții

1. Se consideră un set de date pentru o variabilă X: $(2, 4, 5, 2, 3, 4)$, la nivelul unei colectivități de volum $N = 6$.

Utilizând schema extragerii aleatoare fără revenire, să se extragă toate eșantioanele de volum $n = 3$. Se cere:

- să se determine numărul de eșantioane posibil de extras;
- să se construiască distribuția fiecărui eșantion extras;
- să se calculeze probabilitatea de a extrage un eșantion;
- să se calculeze probabilitatea de includere a unei unități în eșantion;

- să ce construiască distribuția mediei de selecție și să se verifice proprietatea de nedeplasare;
- să se reprezinte grafic distribuția mediilor de eșantion și să se interpreteze rezultatul obținut.

2. Pentru un eșantion de 15 studenți, extrași aleator fără revenire dintr-o populație de 100 de studenți, s-au înregistrat următoarele valori pentru timpul de rezolvare a unui test, exprimat în minute: (45, 50, 46, 40, 38, 42, 40, 45, 50, 44, 42, 48, 41, 42, 46). Se cere:

- să se estimeze punctual media și varianța variabilei pentru populația din care s-a extras eșantionul;
- să se estimeze prin interval de încredere media populației, cu o încredere de 95%;
- dacă se cunoaște dintr-un studiu anterior abaterea standard a variabilei analizate la nivelul întregii populații, având o valoare de 5 minute, să se estimeze prin interval de încredere media populației totale folosind această informație și să se compare cu rezultatul obținut la punctul anterior.

3. Un eșantion de 25 de parcări de mașini este extras aleator fără revenire dintr-o populație de 4000 de parcări ce aparțin unei regiuni geografice. La nivelul eșantionului s-a estimat un număr mediu de mașini parcate zilnic egal cu 120 și o abatere standard de 10 mașini. Se cere:

- să se estimeze punctual și prin interval de încredere numărul mediu de mașini parcate la nivelul întregii populații;
- să se estimeze punctual și prin interval de încredere numărul total de mașini parcate zilnic la nivelul populației totale, pentru o probabilitate de 0,95;
- să se estimeze prin interval de încredere varianța numărului de mașini din parcări la nivelul întregii populații, pentru o încredere de 95%.

4. Un studiu al opiniei politice pe un eșantion de 1000 de persoane arată că 56% dintre cei intervievați sunt de acord cu schimbarea constituției. Să se estimeze prin interval de încredere, cu o probabilitate de 0,97, procentul de persoane, la nivelul populației totale, care sunt de acord cu schimbarea.

5. În vederea realizării unui studiu de opinie la nivelul studenților privind renunțarea la locurile cu taxă, se cere să se estimeze volumul eșantionului care trebuie extras din populație pentru a estima ponderea studenților care doresc această schimbare la nivel universitar cu o eroare maxim admisibilă de 1,7%. Pe acest subiect, pentru populația studenților nu există informații suplimentare, iar probabilitatea cu care se garantează rezultatele este de 97%.

6. Pentru datele de la problema 5 există aceeași cerință, dar se cunoaște volumul populației studenților $N = 9000$. Comparați rezultatul cu cel obținut la

exercițiul 4. Dacă, în plus, se cunoaște o varianță a populației egală cu 0,2, ce se întâmplă cu volumul eșantionului față de celelalte două cazuri?

7. Într-o anchetă se dorește să se estimeze suma medie cheltuită de un student pentru activități culturale într-o lună. Se cere să se calculeze volumul eșantionului necesar unui sondaj aleator fără revenire dintr-o populație de 4000 de studenți, cunoscând că abaterea standard la nivelul populației pentru variabila analizată este de 25 de lei și că se dorește ca estimarea să se realizeze cu o eroare standard de maxim 5 lei.

8. Considerând aceleași condiții de la exercițiul precedent, se cere să se estimeze volumul eșantionului știind că varianța estimatorului mediei nu trebuie să depășească o valoare egală cu 2.

9. S-a construit un sondaj aleator simplu fără revenire pentru care se cunoaște fracția de sondaj f. Pentru care dintre următoarele elemente de calcul este necesar să se utilizeze valoarea lui f?

- estimarea punctuală a mediei populației totale;
- testarea diferenței dintre valoarea mediei populației și o valoare critică cunoscută;
- estimarea prin interval de încredere a proporției;
- calculul erorii maxim admisibile pentru media populației.

10. În estimarea calculului volumului eșantionului, se poate utiliza o condiție cu privire la mărimea intervalului de încredere pentru parametrul medie. În condițiile în care se cunoaște o asemenea limită și se dispune despre o informație cu privire la gradul de omogenitate al populației, se cere să se stabilească o relație de calcul pentru volumul eșantionului în cazul sondajului aleator simplu fără revenire.

11. Similar cu exercițiul precedent, se dispune de o informație privind o limită a coeficientului de variație al estimatorului parametrului medie și se cere să se obțină o relație de calcul pentru volumul eșantionului în cazul unui sondaj aleator cu și fără revenire, cunoscând dintr-un studiu anterior coeficientul de variație la nivelul populației totale.

12. Într-o cercetare prin sondaj se dorește să se estimeze numărul de cărți citite de o persoană în ultimele 6 luni. Ancheta se referă la persoanele cu vârsta de peste 14 ani dintr-o anumită localitate. Se consideră că inițiatorul anchetei este primăria respectivei localități și că aceasta dispune de o listă cu adresele tuturor gospodăriilor ce poate fi utilizată în procesul de selecție a unui eșantion aleator. Se cere:

- să se definească populația de referință supusă cercetării și unitățile statistice individuale precizând distincția între unitățile de eșantionare și cele de observare;
- să se definească baza de sondaj și ce elemente trebuie să conțină pentru a permite selecția;
- să se specifice ce procedeu de selecție aleatoare se poate aplica în acest caz;
- să se discute principalele probleme legate de construirea eșantionului.

13. Prezentați câteva dintre avantajele și limitele sondajului aleator simplu.

14. Dați câteva exemple de situații în care nu se poate aplica metoda sondajului aleator pentru investigarea unui fenomen real.

15. Într-o localitate există 5000 de gospodării. S-a extras aleator fără revenire un eșantion de 100 de unități și s-a observat că doar 25 de gospodării dețin un autovehicul. Se cere să se estimeze punctual și prin interval de încredere numărul de gospodării care dețin un autovehicul la nivelul întregii localități.

16. Discutați posibilitatea realizării unor anchete în care se poate aplica metoda sondajului aleator simplu, iar eșantionul să fie extras după procedeul selecției cu revenire.

Capitolul 3

SONDAJUL NON-ALEATOR (EMPIRIC)

Cercetările prin sondaj sunt condiționate de cost, timp și resurse umane. De asemenea, decizia pentru un anumit tip de sondaj poate avea rațiuni teoretice și metodologice sau cauze practice. În studiile cantitative, miza o reprezintă extrapolarea rezultatelor obținute la nivel de eșantion pentru populația de referință. Ca urmare, condiția de reprezentativitate este esențială, iar sondajul aleator reprezintă cea mai bună soluție. Decizia de a apela la un sondaj empiric se ia la limită, dacă nu există alte soluții. Chiar dacă nu au la bază un principiu de selecție matematic, pentru sondajele empirice se asigură un nivel de reprezentativitate prin utilizarea unor principii de selecție ce respectă demersul unei cercetări științifice. În cazul studiilor calitative, de regulă, obiectivul principal nu este reprezentat de generalizarea rezultatelor la nivelul populației. Asemenea cercetări utilizează cu succes sondajele empirice pentru că servesc scopului principal al unei cercetări în profunzime, pe un eșantion de unități extras dintr-o populație pe baza unor criterii bine întemeiate teoretic. Nu în ultimul rând, sondajele non-aleatoare sunt utilizate cu succes în studii cu obiectiv explorator, în etapele preliminare ale unor cercetări cantitative de amploare sau în studii pilot care pregătesc cercetări exhaustive cum ar fi recensămintele.

Una dintre condițiile de bază cerute de sondajele aleatoare este aceea de a dispune de o bază de sondaj. Deseori, în practică, o asemenea listă cu toate unitățile populației este dificil de obținut sau chiar imposibil de constituit. În asemenea situații, realizarea unei anchete prin sondaj presupune apelul la o serie de criterii de selecție care se bazează pe informația disponibilă și utilizează principii de selecție empirice.

În acest capitol sunt prezentate câteva aspecte specifice tehnicilor de sondaj care au la bază principii de selecție non-aleatoare. Capitolul se deschide cu câteva observații asupra principiilor de selecție în cazul sondajelor empirice. Urmează

prezentarea metodelor cel mai frecvent utilizate în practică. Ne oprim în special la sondajul pe cote și sondajul pe baza unităților tip. În ultima secțiune sunt amintite și alte metode care se utilizează în condiții specifice.

Sondajele non-aleatoare permit construirea unor eșantioane ce respectă condiția de reprezentativitate fără apelul la metode matematice de selecție. Asemenea situații apar în practica sondajelor din mai multe rațiuni. Prima, și cea mai des invocată, este lipsa unei baze de sondaj, condiție expres cerută de sondajele aleatoare. Există însă și alte motive pentru apelul la metode non-aleatoare. Uneori, interesul practic al sondajului nu este legat de realizarea de inferențe statistice, ci doar de a realiza o cunoaștere preliminară a unei probleme, astfel că apelul la metodele probabiliste nu mai este justificat. Studiile pilot care pregătesc alte cercetări complexe se pot realiza cu succes prin apelul la metode de sondaj non-aleatoare. O altă situație des întâlnită este legată de costul anchetei. Sondajul non-aleator implică costuri mult mai mici și chiar un timp de realizare mult redus. Când costul și timpul sunt determinante, sondajul non-aleator reprezintă o soluție convenabilă.

3.1. Principiile selecției non-aleatoare

Sondajele non-aleatoare și-au câștigat locul în rândul metodelor de cercetare parțială, în ciuda limitelor privind posibilitățile de a evalua erorile de reprezentativitate și de a extrapola rezultatele obținute. În cazul sondajelor non-aleatoare, selecția unităților care compun eșantionul nu se realizează fără apelul la un principiu sau un raționament care să asigure proprietatea de reprezentativitate. Asemenea principii de selecție s-au constituit în funcție de natura populației, tema de cercetare, informațiile disponibile, scopul cercetării etc.

Probabil că cel mai cunoscut principiu de selecție este cel al respectării unei structuri a populației totale la nivelul eșantionului pe baza unor caracteristici determinate în raport cu tema de cercetare și natura populației. Ideea este de a construi un eșantion care să reprezinte o *miniatură* a populației, asigurând astfel condiția de reprezentativitate printr-o concordanță de structură. O asemenea abordare presupune cunoașterea structurii populației după anumite criterii semnificativ corelate cu tema de cercetare și posibilitatea selecției unui eșantion care să respecte această structură. Chiar dacă asemenea informații există, pentru a asigura un eșantion reprezentativ, este important să se stabilească precis care sunt cele mai potrivite variabile și care este numărul optim de criterii care să asigure procesul de selecție.

Un principiu des întâlnit este cel al *accesibilității* unităților din populație. În eșantion sunt incluse unitățile care sunt accesibile atât prin posibilitatea efectivă de a răspunde anchetei, cât și prin proximitatea spațială, prin ușurința cu care pot fi incluse în eșantion. În cazul unei anchete în care sunt intervievate persoane fizice,

criteriul accesibilității presupune includerea în eșantion a persoanelor care doresc să participe la anchetă, care răspund pozitiv unei asemenea invitații, fără a aplica alte criterii în plus. Subiecții sunt aleși pentru că sunt disponibili, pentru că sunt ușor de selectat în eșantion. Acest criteriu se poate aplica populațiilor relativ omogene, caz în care reprezentativitatea nu cere o structurare a priori a populației analizate.

Un alt criteriu de selecție este legat de *timp*. Eșantionul se poate constitui prin selecția unităților din populație care pot fi identificate la un moment dat sau într-un anumit interval de timp. Procedeul se repetă de câteva ori până la constituirea întregului eșantion. Rațiunea unei asemenea proceduri este dată de dinamica populației, iar secvențierea în timp a procesului de selecție ar asigura un eșantion care să respecte condiția de reprezentativitate. Prin realizarea selecției la intervale de timp bine specificate, se consideră că procedeul asigură accesul la întreaga diversitate a populației.

Un criteriu similar de selecție poate fi legat de *spațiu*. Acoperirea spațiului pe care se manifestă populația de referință, după un anumit criteriu geografic clar, reprezintă o altă posibilitate de selecție. În general, se realizează un traseu și puncte de selecție care acoperă spațial populația și se asigură astfel o corespondență între eșantion și întreg.

Cele două criterii prezentate anterior, cel temporal și cel spațial, pot fi utilizate împreună în procesul de construire a unui eșantion reprezentativ în contextul unui sondaj empiric. Abordarea beneficiază de avantajele selecției după ambele coordonate.

O altă modalitate de a construi eșantionul este reprezentată de *expertiza* sau judecata de valoare pe care o oferă specialistul care realizează ancheta. Autoritatea epistemică poate funcționa ca un criteriu de selecție și se manifestă, în general, prin definirea unui profil al candidatului care urmează să fie selectat. În acest caz, expertul poate decide să selecteze din populație anumite unități, cum ar fi cazurile tipice, extreme, critice, particulare, comune etc. Decizia pentru o anumită categorie sau pentru un anumit profil al unităților selectate poate depinde de natura populației studiate sau a temei de cercetare.

Principiul *bulgărelui de zăpadă* reprezintă un alt exemplu de raționament după care se construiește eșantionul. În acest caz, identificarea persoanelor incluse în eșantion se realizează cu ajutorul celor selectate deja. Posibilii noi candidați sunt propuși de cei intervievați în măsura în care persoanele deja selectate pot cu ușurință face o evaluare a condițiilor impuse de anchetă și pot identifica asemenea unități. Acest principiu admite și riscul unor erori sistematice care însoțesc practica propunerii de noi unități de către cele ce au participat deja la anchetă. De exemplu, persoanele deja intervievate pot recomanda unități cu caracteristici similare.

Nu în ultimul rând, un procedeu de selecție este acela care lasă decizia de a participa sau nu la anchetă respondentului sau unității individuale care primesc invitația de a fi incluse în eșantion. Această idee are la bază *voluntarismul* și *disponibilitatea* unităților din populație de a se implica activ într-o cercetare pe bază de sondaj. În acest sens, pentru a sublinia importanța participării la anchetă, este semnificativă modalitatea prin care posibilii respondenți sunt invitați să răspundă sondajului, precum și metodele și instrumentele care promovează cercetarea și pot convinge potențialii participanți să se implice.

3.2. Sondajul pe cote

Cea mai cunoscută și poate cel mai des utilizată metodă de sondaj non-aleator poartă numele de sondaj pe cote (Ardilly, 1994, p. 154). Argumentele în favoarea acestei poziții sunt legate în special de calitatea rezultatelor. Deși nu se poate estima sau măsura precizia rezultatelor obținute, sondajele pe cote oferă rezultate apropiate calitativ de sondajele care implică selecția aleatoare într-o anumită etapă a procesului de extragere a eșantionului (sondajele mixte). Metoda este frecvent utilizată în studiile economice și sociale datorită facilității cu care se obține eșantionul, a costurilor reduse și a calității rezultatelor.

În cele ce urmează, se prezintă detaliat principiul de selecție, planul de sondaj, o serie de probleme organizatorice, avantajele și limitele metodei de sondaj pe cote.

3.2.1. Principiul de bază. Planul de sondaj

Metoda sondajului pe cote presupune obținerea unui eșantion ca o miniatură a populației după o structură dată de anumite variabile esențiale pentru anchetă, numite *variabile de control*. La baza metodei stă ipoteza de corelare semnificativă a variabilelor de structurare a populației cu tema sau variabila supusă cercetării în respectiva anchetă. Dacă se respectă această condiție, eșantionul extras este considerat reprezentativ și oferă rezultate de o bună calitate.

Un plan de sondaj pe cote trebuie construit după aceleași reguli metodologice specifice oricărui sondaj statistic. Mai precis, trebuie rezolvate următoarele probleme specifice:

- definirea populației - caracteristica ce definește natura populației, spațiul și timpul;
- definirea informațiilor suplimentare - în acest caz, este vorba despre variabilele de control;
- definirea eșantionului - volumul și structura acestuia, metoda de selecție a unităților din populație;
- stabilirea gradului de precizie și a erorilor de eșantionare.

În această secțiune, pentru sondajul pe cote ne oprim la variabilele de control, definirea eșantionului, erorile de eșantionare și precizia metodei.

a. Alegerea variabilelor de control

Calitatea sondajului pe cote depinde de alegerea variabilelor de control. În această poziție, se pot afla variabilele care:

- au o distribuție cunoscută la nivelul populației totale - restricție destul de importantă, care poate crea probleme practice. Frecvent, pentru anumite variabile, se obțin informații din surse de date oficiale, cum ar fi recensământul populației sau estimații pornind de la aceste date. Există însă variabile de control importante pentru care asemenea date nu sunt disponibile. De asemenea, uneori, informația privind variabilele de control nu este actuală și generează erori de reprezentativitate;
- sunt puternic corelate cu variabilele studiate în anchetă - condiție indispensabilă asigurării calității rezultatelor;
- pot fi ușor observate de operatori în faza de culegere a datelor - operatorii au anumite consemne după care identifică persoanele care fac parte din eșantion. Dacă acest proces nu se poate realiza ușor, pot apărea erori de observare importante;
- permit corectarea erorilor care pot apărea din cauză că anumite categorii de persoane sunt mai ușor accesibile și de recunoscut decât altele (Ghiglione și Matalon, 1985, p. 39).

În faza de elaborare a planului de sondaj, se propune o listă de variabile de control care respectă criteriile precizate mai sus. Utilizând experiența anchetelor anterioare, se pot construi ușor aceste liste cu variabile. De asemenea, se verifică disponibilitatea datelor privind structura populației pentru aceste variabile propuse pentru a îndeplini rolul de variabile de control.

În anchetele socio-economice, rolul acestor variabile îl iau caracteristicile socio-demografice ale persoanelor anchetate: vârsta, genul, starea civilă, nivelul de educație, mediul de proveniență, profesia etc. Asemenea informații pentru o populație există, de regulă, în registrele și statisticile oficiale și permit structurarea populației și calculul cotelor.

b. Stabilirea cotelor și construirea eșantionului

Cotele sunt proporții din populație, stabilite pe baza variabilelor de control prin valorile sau variantele sale. De exemplu, după variabila genul persoanei, într-o populație pot exista 52% femei și 48% bărbați. Cele două procente reprezintă cote, ponderi care vor fi respectate și la nivelul eșantionului. Un plan de sondaj eficient cere o extragere a eșantionului după un număr echilibrat de variabile de control: nu prea mare, pentru că ridică atât costurile, cât și nivelul de dificultate al selecției unităților; nu prea mic, pentru că reduce reprezentativitatea eșantionului și precizia.

Cotele pot fi utilizate fie independent, după fiecare variabilă, fie încrucişat. Cotele încrucişate ridică probleme atât prin constrângerile date de existenţa informaţiilor pentru determinarea acestora, cât şi legate de extragerea eşantionului. Uneori, pentru un operator, a identifica o persoană după cote încrucişate este aproape imposibil, iar timpul de realizare a observării creşte considerabil, riscul de a apărea erori fiind mult mai mare. În practică, se recomandă încrucişarea cotelor doar dacă este cu adevărat necesară. Procedura se realizează după două variabile de control mai importante, iar pentru celelalte variabile se stabilesc cote marginale.

O problemă importantă pe care o ridică construirea eşantionului într-un sondaj pe cote este volumul acestuia. Fiind un sondaj empiric, nu există o bază teoretică şi metodologică de calcul al volumului eşantionului. Soluţia este la aprecierea specialistului: fie să se utilizeze un volum de eşantion dintr-o metodă cu care acest tip de sondaj se poate compara, fie să se aprecieze în funcţie de mărimea şi omogenitatea populaţiei. În practică, se utilizează frecvent un volum al eşantionului în jur de 900-1000 de unităţi (sau un volum apropiat de sondajul aleator simplu). Decizia asupra volumului eşantionului îi revine expertului responsabil cu realizarea planului de sondaj şi aceasta poate depinde de o serie de informaţii care provin din studii anterioare sau care vizează natura şi omogenitatea populaţiei.

Extragerea unităţilor din eşantion presupune mai întâi constituirea unei structuri la nivelul eşantionului, utilizând cotele de la nivelul populaţiei totale. Dacă se utilizează un eşantion de 1000 de unităţi, atunci structura eşantionului este facil de obţinut pe baza ponderilor de la nivelul populaţiei totale.

Exemplu

La nivelul României, pe lângă variabilele vârstă şi gen, o structurare facilă a populaţiei se poate face după următoarele criterii: regiunea de dezvoltare, judeţ, regiunea istorică, mediul de proveniență.

Tabelul 3.1. *Structura populaţiei României pe regiuni la recensământul din 2011*

Regiunea	Persoane	Cote
Nord-Vest	2600132	12,92
Centru	2360805	11,73
Nord-Est	3302217	16,41
Sud-Est	2545923	12,65
Sud-Muntenia	3136446	15,59
Bucureşti-Ilfov	2272163	11,29
Sud-Vest Oltenia	2075642	10,32
Vest	1828313	9,09
Total	**20121641**	**100**

Sursa: Realizat de autor pe baza datelor oferite de INS, 2018

Pentru acest exemplu, un eșantion de 1000 de persoane extrase din populație va avea structura din tabelul de mai jos.

Tabelul 3.2. *Structura unui eșantion reprezentativ pe regiuni de dezvoltare*

Regiunea	Cote	Persoane
Nord-Vest	12,92	129
Centru	11,73	117
Nord-Est	16,41	164
Sud-Est	12,65	127
Sud-Muntenia	15,59	156
București-Ilfov	11,29	113
Sud-Vest Oltenia	10,32	103
Vest	9,09	91
Total	**100**	**1000**

Sursa: Realizat de autor pe baza datelor oferite de INS, 2018

Tabelul 3.3. *Structura populației României pe regiuni de dezvoltare și medii la recensământul din 2011*

Regiunea	Persoane		Pondere	Persoane în eșantion
Nord - Vest	Total	2600132	12,92	129
	Rural	*1233182*	*47,4*	*61*
	Urban	*1366950*	*52,6*	*68*
Centru	Total	2360805	11,73	117
	Rural	*992497*	*42*	*49*
	Urban	*1368308*	*58*	*68*
Nord - Est	Total	3302217	16,41	164
	Rural	*1927423*	*58,4*	*96*
	Urban	*1374794*	*41,6*	*68*
Sud - Est	Total	2545923	12,65	127
	Rural	*1183912*	*46,5*	*59*
	Urban	*1362011*	*53,5*	*68*
Sud - Muntenia	Total	3136446	15,59	156
	Rural	*1893565*	*60,4*	*94*
	Urban	*1242881*	*39,6*	*62*
București - Ilfov	Total	2272163	11,29	113
	Rural	*221710*	*9,8*	*11*
	Urban	*2050453*	*90,2*	*102*
Sud - Vest Oltenia	Total	2075642	10,32	103
	Rural	*1117664*	*53,8*	*55*
	Urban	*957978*	*46,2*	*48*
Vest	Total	1828313	9,09	91
	Rural	*692898*	*37,9*	*34*
	Urban	*1135415*	*62,1*	*57*
Total România	**Total**	**20121641**	**100**	**1000**
	Rural	*9262851*	*45,9*	*459*
	Urban	*10858790*	*54,1*	*541*

Sursa: Realizat de autor pe baza datelor oferite de INS, 2018

Tabelul 3.4. *Structura populației României pe grupe de vârstă și pe gen, 2016*

Grupa de vârstă	Persoane		Pondere	Eșantion
0-4 ani	Total	953912	4,83	48
	Masculin	*490819*	*51,45*	*25*
	Feminin	*463093*	*48,55*	*23*
5-9 ani	Total	1061963	5,37	54
	Masculin	*544904*	*51,31*	*28*
	Feminin	*517059*	*48,69*	*26*
10-14 ani	Total	1049118	5,31	53
	Masculin	*539643*	*51,44*	*27*
	Feminin	*509475*	*48,56*	*26*
15-19 ani	Total	1083079	5,48	55
	Masculin	*554837*	*51,23*	*28*
	Feminin	*528242*	*48,77*	*27*
20-24 de ani	Total	1079644	5,46	55
	Masculin	*558942*	*51,77*	*28*
	Feminin	*520702*	*48,23*	*27*
25-29 de ani	Total	1357286	6,87	69
	Masculin	*701855*	*51,71*	*36*
	Feminin	*655431*	*48,29*	*33*
30-34 de ani	Total	1305682	6,61	66
	Masculin	*671880*	*51,46*	*34*
	Feminin	*633802*	*48,54*	*32*
35-39 de ani	Total	1539105	7,79	78
	Masculin	*783317*	*50,89*	*40*
	Feminin	*755788*	*49,11*	*38*
40-44 de ani	Total	1515385	7,67	77
	Masculin	*777690*	*51,32*	*40*
	Feminin	*737695*	*48,68*	*37*
45-49 de ani	Total	1635199	8,28	83
	Masculin	*838381*	*51,27*	*43*
	Feminin	*796818*	*48,73*	*40*
50-54 de ani	Total	1071973	5,42	54
	Masculin	*540352*	*50,41*	*27*
	Feminin	*531621*	*49,59*	*27*
55-59 de ani	Total	1326453	6,71	67
	Masculin	*640527*	*48,29*	*32*
	Feminin	*685926*	*51,71*	*35*
60-64 de ani	Total	1345095	6,81	68
	Masculin	*621556*	*46,21*	*31*
	Feminin	*723539*	*53,79*	*37*
65-69 de ani	Total	1103942	5,59	56
	Masculin	*489948*	*44,38*	*25*
	Feminin	*613994*	*55,62*	*31*
70 de ani și peste	Total	2332478	11,80	118
	Masculin	*895160*	*38,38*	*45*
	Feminin	*1437318*	*61,62*	*73*
Total România	**Total**	**19760314**	**100**	**1000**
	Masculin	*9649811*	*48,83*	*488*
	Feminin	*10110503*	*51,17*	*512*

Sursa: Realizat de autor pe baza datelor oferite de INS, 2018

În cazul în care la variabilele de control adăugăm mediul de proveniență, se poate construi o structură a eșantionului prin încrucișarea celor două criterii, așa cum se prezintă informațiile din tabelul 3.3.

Dacă într-o anchetă este importantă grupa de vârstă, distinct sau împreună cu alte variabile de control, se utilizează fie datele statistice de la recensământ, fie datele curente estimate. De exemplu, conform datelor statistice oferite de INS, pe anul 2016, structura populației României pe grupe de vârstă și pe gen este prezentată în tabelul 3.4. Cotele care se pot utiliza într-un sondaj sunt prezentate pe ultima coloană.

c. Erorile de eșantionare

O primă eroare care poate apărea într-un sondaj pe cote este eroarea de selecție sau de includere în eșantion (Ardilly, 1994, p. 155). Această eroare este cauzată în principal de alegerea variabilelor de control. Dacă lipsesc anumite variabile de control sau nu au fost bine alese, apar erori care afectează direct calitatea rezultatelor. Această categorie de erori poate fi diminuată atât prin alegerea judicioasă a variabilelor de control, cât și prin stabilirea clară a consemnelor pentru operatori.

De exemplu, într-o anchetă de opinie cu privire la viața politică, dacă se consideră ca variabile de control doar vârsta și genul, poate să apară eroarea de selecție prin introducerea în eșantion în mod preponderent a persoanelor de o anumită categorie socio-profesională, variabilă care este determinantă pentru opinie.

O altă sursă de eroare de includere în eșantion o constituie actualitatea informației cu privire la variabilele de control. În cazul în care cotele utilizate în eșantionare nu mai sunt actuale, rezultatul obținut nu va mai fi „reprezentativ" pentru populația studiată.

De exemplu, se consideră o variabilă de interes Y. La nivelul eșantionului se calculează o medie după relația:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{k} \bar{y}_i n_i = \sum_{i=1}^{k} \bar{y}_i \frac{n_i}{n} = \sum_{i=1}^{k} \bar{y}_i \frac{N_i}{N}$$

Dacă însă cotele N_i/N nu sunt actuale, media de la nivelul eșantionului va „estima" deplasat media de la nivelul populației totale.

d. Precizia rezultatelor

Deoarece nu se cunosc probabilitățile de includere în eșantion, nu se pot calcula erorile de eșantionare. Cu toate acestea, condițiile de elaborare a unui plan de sondaj pe cote oferă posibilitatea aprecierii erorilor apărute în urma eșantionării. Se pot face câteva observații în acest sens.

Deşi nu poate fi calculată, se poate presupune că varianţa unui sondaj pe cote este mică, adică rezultatele sunt precise. Argumentul acestei afirmaţii derivă din faptul că variabilele de interes sunt puternic corelate cu variabilele de control (pentru acestea din urmă varianţa este nulă).

În cazul unui sondaj aleator de volum mic ($n < 1000$), varianţa estimatorului este semnificativă, fiind invers proporţională cu volumul eşantionului. Pentru un eşantion de acelaşi volum, un sondaj pe cote poate oferi un rezultat mai precis, cu o varianţă mai scăzută, deşi rezultatul poate fi deplasat prin eroarea de selecţie. Astfel, pentru eşantioane mici, poate fi preferat un sondaj pe cote, în schimb, pentru eşantioane mari, sondajul probabilist dă rezultate mai bune, iar precizia poate fi măsurată.

3.2.2. Probleme organizatorice specifice

În organizarea unui sondaj pe cote, pentru a obține rezultate de calitate, este esențial să se realizeze o bună selecție a operatorilor, să se constituie fișe de operatori analitice care să specifice modul de identificare a unităților care vor fi selectate și să se asigure un control permanent al operatorilor în timpul culegerii datelor.

a. Operatorii și metodele de administrare

Există două metode de administrare a chestionarului: utilizarea operatorilor care să completeze răspunsurile în timpul unui interviu și autoadministrarea sau completarea chestionarului direct de către respondent.

În cazul sondajelor pe cote administrate de operatori, pe lângă culegerea efectivă a datelor, operatorilor le revine și sarcina identificării unităților care participă la selecție. Din această pricină, această metodă de sondaj impune o atenție aparte etapei de selecție și instruire a operatorilor.

Sondajele pe cote se pot realiza ocazional, pentru o anumită cercetare, sau pot fi organizate regulat, caz în care se pot utiliza operatori care au o anumită experiență și care cunosc detaliile anchetei. În primul caz, efortul este mai important, pentru că selecția operatorilor necesită un timp de lucru mai mare și, implicit, există riscuri de a avea erori mai mari. În al doilea caz, se constituie o echipă de operatori care participă constant la anchete, iar efortul selecției și al instruirii se realizează o singură dată.

Pentru selecția operatorilor este important să se ia în calcul aria geografică a anchetei și, pe cât posibil, să se selecteze persoane care să aibă domiciliul în apropierea zonei de culegere a datelor. Ideea este de a reduce cheltuielile de transport și cazare, pe de o parte, dar și de a facilita procesul de identificare a unităților de eșantionare, pe de altă parte.

Recrutarea operatorilor a devenit deja un subiect binecunoscut. Firmele de sondaje își prezintă ofertele pe site-urile de specialitate și anunță caracteristicile

candidaților și fișa postului. De asemenea, firmele de sondaj au elaborate manuale pentru operatori în care sunt prezentate în detaliu cerințe pentru ocuparea unei asemenea poziții, dar și elemente de instruire care să ajute la culegerea datelor.

Exemplu

Pe site-ul ejobs (www.ejobs.ro) apare un anunț pentru poziția de operator de sondaj de opinie. Conform acestei surse, candidatul ideal este persoana:

- *organizată;*
- *rezistentă la munca de teren;*
- *atentă la detalii;*
- *meticuloasă;*
- *disponibilă pentru realizarea de proiecte periodice;*
- *cu reale abilități de comunicare;*
- *cu putere de convingere și tenacitate.*

La descrierea acestui post de muncă, apar următoarele caracteristici:

- *completarea chestionarelor pe teren cu respectarea strictă a metodologiei;*
- *îndeplinirea normei de lucru;*
- *respectarea programului;*
- *seriozitatea;*
- *raportarea periodică.*

b. Respectarea cotelor. Fișe de operatori

Fișele de operator reprezintă instrumentul de bază pe care îl utilizează cei care administrează chestionarele pe teren. Aceste fișe sunt individualizate pentru fiecare operator și prezintă numărul și structura persoanelor care trebuie intervievate, precum și o serie de informații suplimentare care să ajute la identificarea acestora.

De regulă, fișele de operatori conțin informații pe baza cotelor marginale, calculate pentru fiecare variabilă de control. În situația în care se combină criterii care nu conduc la un număr prea mare de combinații și la dificultăți de selecție, se pot constitui și asemenea fișe cu cote încrucișate. Evident, pentru cotele încrucișate trebuie să existe informația disponibilă la nivelul populației (în practică, asemenea informații nu sunt frecvente).

Exemplu

Cel mai simplu exemplu de fișă de operator cu cote încrucișate este fișa cu două variabile de control și cu două valori pentru fiecare variabilă. De exemplu, dacă într-un studiu se utilizează variabilele gen și mediu de proveniență, iar un operator are de intervievat 10 persoane, se poate realiza o fișă de operator conform schemei prezentate în tabelul de mai jos.

Tabelul 3.5. *Model fișă de operator*

	Urban	Rural	**Total**
Masculin	2	2	4
Feminin	3	3	6
Total	5	5	10

c. Controlul operatorilor

În cazul sondajului pe cote, controlul operatorilor este un procedeu dificil de realizat, în comparație cu sondajele aleatoare, unde se dispune de o informație privind identificarea unităților selectate în eșantion. Pentru sondajul empiric, operatorul are o anumită libertate de alegere a unităților, cu restricția de a respecta cotele și informațiile suplimentare privind zona geografică sau perioada de timp.

Totuși controlul operatorilor este important, pentru a evita apariția unor erori semnificative de culegere a datelor. Din pricina dificultății, o parte a acestei etape este transferată etapei de elaborare a fișelor de operator și a instruirii. Fișele pot conține informații suplimentare și restricții care să ghideze operatorul în selecția unităților, pentru a evita anumite erori predictibile (cum ar fi selecția unor grupuri de persoane din aceeași arie geografică sau dintr-un singur interval de timp).

3.2.3. Avantaje și limite

Sondajul pe cote are atât adepți, cât și contestatari fervenți. Este o metodă care are o anumită suplețe și este relativ ușor de realizat, dar admite și o serie de limite care trebuie cunoscute și asumate.

a. Avantaje

Metoda sondajului pe cote este cunoscută în primul rând prin calitatea rezultatelor, pentru că reprezentativitatea eșantionului este asigurată prin corespondența de structură cu populația de referință pe baza variabilelor de control.

În comparație cu sondajul aleator, avantajele sondajul pe cote sunt legate de:

- dimensiunea economică - acest tip de sondaj este mai ieftin. Costurile sunt mai reduse pentru că eșantionul este mai mic, cheltuielile de deplasare sunt substanțial reduse, iar informația necesară este fie gratuită, fie mult mai ieftină, comparativ cu informația necesară realizării sondajului aleator simplu;
- timpul de realizare - ancheta se efectuează mai rapid. Toate etapele de realizare a anchetei presupun un timp mai scurt, dar în special cel de culegere a datelor;
- posibilitățile reale de realizare - nu presupune o bază de sondaj, ceea ce, în practică, deseori este imposibil de obținut, astfel că sondajul pe cote poate fi singura metodă accesibilă și posibil de aplicat. Acesta reprezintă un avantaj decisiv în practica sondajelor (Desabie, 1966, p. 54);

- precizie - printr-o muncă analitică și atentă de selecție a eșantionului, se pot obține rezultate de calitate bună, comparabile cu cele ale sondajelor mixte;
- tratarea non-răspunsurilor - dacă o persoană refuză să participe la anchetă, aceasta poate fi înlocuită cu una care prezintă caracteristici similare, ceea ce nu se poate face în cazul sondajului aleator (caz în care se mărește din start volumul eșantionului pentru a acoperi asemenea situații).

b. Limite

Cea mai serioasă critică adusă sondajului pe cote este lipsa de fundament teoretic, baza empirică de selecție a eșantionului. O asemenea limită se prelungește în imposibilitatea de a evalua sau de a măsura gradul de precizie a rezultatelor. Pentru rezultatele obținute cu această metodă de sondaj nu se pot calcula intervale de încredere, nu se poate determina mărimea erorilor și nu se pot realiza teste statistice.

De asemenea, nu se poate calcula volumul eșantionului și nu există criterii solide care să susțină alegerea unui anumit număr de unități în situații foarte diferite. Utilizarea unui număr de aproximativ 1000 de persoane este o soluție, dar poate crea probleme în cazul populațiilor fie de volum mic (când eșantionul poate fi o pondere însemnată din populație), fie de volum foarte mare.

Fără o rigoare pe măsură, sondajul pe cote poate suferi din pricina modalității de selecție a unităților. În acest caz, pot apărea erori sistematice severe, iar reprezentativitatea poate avea de suferit dacă nu s-au utilizat cele mai bune variabile de control. De exemplu, operatorii pot acoperi doar o anumită arie geografică corespunzătoare populației de referință sau pot intervieva sistematic persoanele care li se par mai adecvate anchetei etc. Deși asemenea erori pot fi limitate printr-o atentă instruire și un control al procesului de culegere a datelor, acestea nu pot fi complet eliminate și nici suficient evaluate.

Nu în ultimul rând, sondajul pe cote admite constrângeri legate de variabilele de control (Desabie, 1966, p. 57). Pe de o parte, este vorba despre dificultatea de a obține informația disponibilă pentru aceste variabile la nivelul populației totale. Pe de altă parte, odată utilizate aceste variabile pentru structurarea populației și apoi a eșantionului, este imposibil să fie introduse în analiza propriu-zisă. Asemenea variabile însă pot fi esențiale pentru analiza temei de cercetare.

3.2.4. Ameliorarea calității unui sondaj pe cote

Ameliorarea calității unui sondaj se poate realiza fie în etapa elaborării planului de sondaj, fie după realizarea anchetei. Analiza posibilităților de îmbunătățire a calității reprezintă un obiectiv esențial al etapei de pregătire și organizare a unui sondaj. Cu această ocazie, se pot identifica anumite posibilități de intervenție asupra etapelor de realizare, odată ce s-a decis utilizarea unei anumite metode. Nu în ultimul rând, analizând limitele și avantajele unui tip de sondaj, se

poate decide asupra unor posibilități de creștere a calității rezultatelor prin anumite modificări la nivelul planului de sondaj sau al anumitor elemente privind costurile și timpul de realizare.

Odată realizată opțiunea pentru o anumită metodă de cercetare selectivă, urmează concentrarea eforturilor pe găsirea posibilităților de creștere a eficienței metodei alese. Planul de sondaj poate fi ameliorat atât prin reducerea erorilor cu care se obțin rezultatele, cât și prin reducerea costurilor și a timpului de realizare a anchetei. În cazul sondajului pe cote, ameliorarea calității se poate face prin eficientizarea muncii la toate etapele realizării anchetei, dar și prin apelul la o metodă mixtă care să rezulte din combinarea sondajului pe cote cu metoda aleatoare.

Erorile de sondaj pot fi limitate la nivelul fiecărei etape de realizare a sondajului: definirea problemei, construirea chestionarului, alegerea variabilelor de control, selecția unităților, culegerea datelor, analiza statistică. Dacă există posibilitatea, se pot aduce elemente complementare ce pot crește precizia rezultatelor și pot reduce costurile.

Un exemplu de ameliorare a sondajului pe cote îl reprezintă realizarea unui sondaj în două etape. Ideea este să se utilizeze informația disponibilă pentru a crea o bază de sondaj cu unități complexe din populație obținute cu ajutorul variabilelor de control. În prima treaptă, se apelează la metoda aleatoare: se extrag aleator asemenea unități complexe din populația de referință. În a doua treaptă, se realizează o extragere pe cote din fiecare unitate selectată aleator.

Avantajele unei astfel de metode sunt:

- realizarea unei bune precizii prin utilizarea unei extrageri aleatoare în prima treaptă;
- reducerea costurilor de constituire a bazelor de sondaj: se construiește doar o bază de sondaj cu unități complexe;
- reducerea costurilor de observare prin utilizarea unei eșantionări pe cote în a doua treaptă.

3.3. Metoda unităților tip

Literatura de specialitate apreciază că aceasta este cea mai empirică sau cea mai slab fundamentată dintre metodele non-aleatoare (Ardilly, 1994, p. 163). Metoda unităților tip are ca principiu de bază împărțirea populației în categorii omogene cu ajutorul unor variabile de control și selectarea în eșantion a unităților care sunt cel mai aproape de media fiecărei categorii sau sub-populații (Desabie, 1966, p. 44). Astfel, se admite ipoteza că unitățile tip se vor comporta similar și în cazul variabilelor de interes studiate în anchetă.

Metoda este foarte utilizată în anchetele exploratorii, ca cercetări ce pregătesc sondaje de volum mare și obținute cu ajutorul metodelor probabiliste sau

mixte. Avantajul acestei metode este acela că oferă informații în timp foarte scurt și cu costuri reduse, iar calitatea rezultatelor este una destul de bună. Limita cea mai importantă a metodei este aceea că nu surprinde variabilitatea la nivelul populației, iar în selecția unităților tip pot apărea erori sistematice. De exemplu, o unitate individuală care se află aproape de medie pentru o anumită variabilă nu este în mod necesar în aceeași poziție în cazul altor variabile. Reprezentativitatea eșantionului este legată de capacitatea cercetătorului de a defini caracteristicile unității tip și apoi a operatorilor de a identifica aceste unități în momentul culegerii datelor.

Domeniul în care este cel mai frecvent aplicată această metodă este cel agricol sau în analiza unor probleme specifice mediului rural. În aceste situații, este mult mai ușor de a identifica unitățile tip, care pot fi gospodării, ferme sau arii de teren agricol bine delimitate.

Ca și în cazul celorlalte tehnici de sondaj empirice, metoda unităților tip este utilă în cazul studiilor exploratorii, pe eșantioane de volum mic, ce se pot realiza într-un timp relativ scurt și cu costuri reduse.

3.4. Alte metode

Cercetarea pe bază de sondaj statistic, deși atât de cunoscută și larg folosită în prezent în toate științele, are o istorie recentă. Acceptarea relativ târzie a sondajului ca o metodă de cunoaștere științifică se datorează unor condiții sociale, politice, științifice și filosofice care s-au manifestat în partea a doua a secolului al XIX-lea și începutul secolului al XX-lea. Înainte de a prezenta pe scurt câteva dintre aceste condiții, este important să ne oprim puțin asupra termenului care dă numele acestei metode.

3.4.1. Metoda voluntariatului

Această metodă de sondaj are la bază principiul autoselecției, adică persoanele din populație decid dacă să participe sau nu la anchetă în urma unei invitații. Așadar, voluntariatul reprezintă criteriul sau principiul care stă la baza acestei metode de selecție. În cazul acestui tip de sondaj, asupra reprezentativității eșantionului pot exista din start cele mai serioase îndoieli. Cele mai cunoscute modalități de organizare a acestor anchete sunt cele care apelează la mass-media (chestionare publicate în ziare sau reviste), la telefonie sau la mediul online (chestionare pe diverse pagini web), la clienții unor firme (de exemplu, degustarea unor produse în magazin) sau la membrii unor asociații nonprofit.

Metoda admite o serie de avantaje care derivă chiar din natura principiului de selecție: costul anchetei este scăzut; nu există criterii predefinite în selecția unităților din populație; datele se culeg relativ rapid, mai ales dacă se face apel la structurile sociale la care participă unitățile din populație (familie, prieteni, membri într-o organizație).

Această metodă admite o limită importantă legată de caracteristicile populației de referință. Voluntariatul reprezintă o trăsătură a unui tip de societate cu o anumită maturitate civică, cu o anumită disponibilitate de a participa la viața publică. Dacă o asemenea cultură este slab dezvoltată, metoda sondajului bazată pe voluntariat poate avea serioase erori de reprezentativitate. Chiar și în condiții în care dispunem de o cultură a voluntariatului, riscul unor erori importante există, pentru că este imposibil de apreciat ponderea și caracteristicile celor care vor răspunde efectiv la anchetă, precum și rata non-răspunsurilor. Pentru o evaluare a calității rezultatelor, este important să se încerce să se obțină informații despre profilul persoanelor care nu au participat la anchetă, al unităților cu o înclinație mai redusă pentru participarea voluntară la asemenea anchete sau la alte acțiuni cu caracter civic.

3.4.2. Metoda selecției bazate pe accesibilitate

Ideea de bază a acestei metode este aceea de a identifica o modalitate facilă de a avea acces la populația de referință. Principiul de selecție se referă la utilizarea contextelor în care un număr mare din populația supusă analizei se află într-un anumit loc și la un moment dat și pot fi ușor de abordat în vederea culegerii datelor. Vorbim despre locuri publice în care se găsesc suficient de multe persoane și care pot fi intervievate: parcuri, magazine, muzee, gări și aeroporturi etc.

Pentru selecția unităților în eșantion, pe lângă condițiile de spațiu și disponibilitate a persoanelor potențiale, se poate aplica și o regulă de selecție. Un exemplu îl constituie regula aplicată în schema extragerii cu pas de numărare: operatorul selectează persoanele cu o anumită frecvență, iar dacă acestea sunt de acord, se realizează culegerea datelor. În plus, pentru a îmbunătăți reprezentativitatea eșantionului, selecția se poate realiza la diferite intervale de timp pe parcursul unei zile sau chiar pe parcursul unei săptămâni. În ciuda unor asemenea măsuri de precauție privind reprezentativitatea, prin natura ei, această metodă de sondaj poate admite importante erori de selecție. Principala obiecție adusă acestei idei de eșantionare este legată de faptul că este dificil de evaluat dacă acea parte a populației de referință accesibilă într-un anumit perimetru este reprezentativă pentru întreaga populație.

Această metodă poate ridica probleme la capitolul non-răspunsuri (persoanele aflate într-un spațiu public au deja un scop sau un interes, iar solicitarea de a participa la un sondaj poate fi privită cu refuz sau neîncredere), dar și la cel al reprezentativității (dacă nu se ține cont de variabilitatea fluxurilor de persoane în locurile unde se realizează selecția).

3.4.3. Metoda itinerariilor

Metoda itinerariilor este o variantă a metodei cotelor. Operatorilor li se impune mult mai strict modalitatea de alegere a unităţilor care participă la anchetă, pe baza unui itinerar (un ansamblu de adrese, străzi sau zone din care se extrag persoane după anumite consemne clare). Această metodă se bazează pe o reprezentativitate spaţială a eşantionului.

Operatorii primesc consemne clare pe o hartă a teritoriului în care se desfăşoară ancheta, iar punctele de selecţie sunt marcate şi identificate pe baza unei adrese. Dacă în anchetă este vizată o localitate, atunci coordonatele principale pentru identificarea unităţilor supuse observării sunt date de adrese, aşa cum sunt definite în documentele civile şi administrative. Operatorii pot dispune de o listă cu clădiri sau cu adrese care să indice punctele de selecţie.

Această metodă are avantajul că lasă mai puţină iniţiativă operatorilor (riscurile apariţiei erorilor de selecţie sunt mai mici), iar procesul de control al culegerii datelor este unul mai facil. De asemenea, această metodă impune utilizarea unor informaţii suplimentare despre populaţia de referinţă, o limită care uneori este dificil de depăşit. Uneori, asemenea anchete pot avea costuri mai ridicate decât alte metode de sondaj empirice, îndeosebi datorită costurilor ridicate cu deplasarea în teritoriu şi acoperirea întregii arii geografice.

3.5. Exerciţii

1. Pentru populaţia studenţilor din oraşul Iaşi se organizează o anchetă prin sondaj pentru a studia opinia despre situaţia căminelor din campus. Se cunosc informaţii despre structura studenţilor pe universităţi, facultăţi şi ani de studii. Să se prezinte argumente pentru ce metodă de sondaj empiric se poate opta în acest caz, având în vedere şi criteriul eficienţei sub raportul costului şi al timpului de realizare a anchetei.

2. Se consideră aceeaşi populaţie şi aceeaşi temă de cercetare de la exerciţiul anterior, însă se presupune că nu se poate dispune de informaţiile legate de structura studenţilor pe universităţi, facultăţi şi ani de studii. Ce abordări se pot utiliza în acest caz pentru realizarea unui sondaj?

3. La nivelul regiunii Nord-Est, se doreşte realizarea unei anchete privind activitatea agricolă a gospodăriilor din mediul rural. Dacă se apelează la un tip de sondaj empiric, care ar fi cel mai adecvat acestei situaţii? Motivaţi răspunsul.

4. Ce tip de sondaj empiric se poate aplica cel mai eficient în cazul studiului opiniei clienţilor unui supermarket privind calitatea unui tip de produse (de exemplu, a produselor lactate de o anumită marcă)?

5. Prezentați pe scurt trei principii de selecție în cazul sondajelor non-aleatoare.

6. Pentru datele disponibile în tabelul 3.6, să se construiască structura unui eșantion de volum $n = 1000$ extras pe cote, utilizând variabila de control *Cartierul*. De asemenea, să se construiască structura unui eșantion similar utilizând simultan variabilele *Cartier* și *Vârstă* și folosind aceeași sursă de informații cu privire la structura populației orașului pe grupe de vârstă.

Tabelul 3.6. *Structura populației municipiului Iași pe cartiere și pe sexe la Recensământul populației din 1992*

Nr. crt.	Cartier	Persoane masculine		Persoane feminine		Total pers.	
		nr.	%	nr.	%	nr.	%
1	Alexandru cel Bun	29963		31295		61258	
2	Bucium	2975		3164		6139	
3	Centru	13747		15188		28935	
4	Copou	8126		7749		15875	
5	Frumoasa	3729		3600		7329	
6	Galata	12271		12636		24907	
7	Moara de vânt	2983		3373		6356	
8	Nicolina CUG	21673		22862		44535	
9	Obreja	2056		2056		4112	
10	Păcurari	19051		19995		39046	
11	Socola-Nicolina	11859		13051		24910	
12	Studențesc	5322		5385		10707	
13	Tătărași	16512		18976		35488	
14	Țicău-Sărărie	4248		6543		10791	
15	Zona ind. Păcurari	462		303		765	
16	Zona ind. Țuțora	11852		11274		23126	
	Total Iași	**166829**		**177450**		**344279**	
	pondere	**%**		**%**		**100%**	

Sursa: *Fișa localității*, Direcția Județeană de Statistică Iași

7. Considerând ancheta prezentată anterior, se cere să se elaboreze un model de fișă de operator care să ia în considerare:

- variabilele *Cartier* și *Gen*;
- variabilele *Cartier*, *Vârstă* și *Gen*.

8. Utilizând criteriile de alegere a unei metode de sondaj prezentate în primul capitol al acestei lucrări, comparați trei dintre cele mai importante metode de sondaj non-aleatoare, în contextul realizării unui sondaj cu caracter explorator cu privire la efectele migrației externe a persoanelor din orașul Iași asupra familiilor din care provin.

9. Pentru sondajul de la exercițiul 6, prezentați câteva exemple de posibilități de creștere a eficienței acestei metode (creșterea preciziei, reducerea costurilor și a timpului de realizare a anchetei). În cazul aceluiași exemplu, comentați posibilitățile de diminuare a volumului eșantionului.

10. Prezentați câteva avantaje și câteva limite ale aplicării sondajului empiric în contextul studiului opiniei politice la nivelul populației României.

11. În cazul unui sondaj de opinie privind activitatea administrațiilor locale din mediul rural din regiunea Nord-Est, dacă se aplică metoda sondajului unităților tip, prezentați o variantă de plan de sondaj pentru realizarea anchetei.

Capitolul 4

SONDAJE MIXTE

Planurile de sondaj mixte[1] presupun utilizarea a două criterii de selecție de natură diferită în procesul de extragere a unui eșantion din populația de referință. Ideea sondajelor mixte este legată de posibilitatea de a îmbunătăți calitatea sondajelor clasice, aleatoare și empirice, prin crearea unor planuri de sondaj care combină elemente metodologice specifice celor două tipuri de anchete. Rezultatul este reprezentat de planurile de sondaj mixte. În acest capitol ne oprim asupra a trei dintre cele mai utilizate metode de sondaj mixte: sondajul stratificat, sondajul în două sau mai multe trepte și sondajul pe serii (cluster).

4.1. Justificarea metodelor mixte

Combinarea celor două procedee clasice de selecție într-un plan de sondaj mixt se bazează pe principiul eficienței activității de cercetare. În primul capitol al acestei lucrări au fost prezentate câteva criterii care pot fi utilizate în decizia sau opțiunea pentru o anumită metodă de sondaje. Aceste criterii sunt: informația disponibilă, precizia rezultatelor, costul, timpul de realizare a anchetei, natura populației de referință, volumul eșantionului, tema de cercetare. În cazul unei anchete concrete, aceste criterii capătă importanță diferită, iar sarcina cercetătorului este aceea de a găsi această ordine de importanță și de a lua o decizie în consecință. Dacă nu există condiții predeterminate (în special de timp și de cost), decizia pentru un anumit plan de sondaj presupune identificarea unei soluții ca urmare a analizei detaliate a criteriilor menționate mai sus.

---

[1] În literatura de specialitate, sintagma *sondaje mixte* este folosită cu două accepțiuni. Prima vizează planurile de sondaj mixte, adică se referă la ideea de a construi un plan de sondaj ce se bazează pe două principii de selecție diferite ce se aplică după un demers bine stabilit. A doua vizează utilizarea celor două tipuri de abordări metodologice clasice, cantitativă și calitativă, într-o cercetare prin sondaj. În această lucrare ne referim la sondajele mixte în prima accepțiune, adică la planuri de sondaj mixte.

Metodele de sondaj clasice au o serie de limite (discutate în capitolele anterioare) care pot fi îndepărtate sau ajustate prin combinarea elementelor metodologice specifice fiecărei metode în parte. Sondajele mixte au la bază ideea de a crește eficiența sondajelor clasice, aleatoare și empirice. Eficiența este înțeleasă în principal prin prisma rezultatelor obținute, însă sunt luate în calcul și timpul și costul anchetei. Pentru sondajele care combină cele două principii de selecție, se utilizează o serie de criterii de analiză pentru a decide asupra celui mai bun plan de sondaj.

Informația suplimentară este cerută de sondajele empirice, cele aleatoare nu necesită o asemenea limită. Însă, dacă o asemenea informație este accesibilă, poate fi folosită împreună cu principiul selecției aleatoare pentru a obține un sondaj mai eficient. Sondajele aleatoare sunt, în general, de volum mare și necesită un timp și un cost mai ridicat. Prin utilizarea informației suplimentare, de regulă, volumul eșantionului se poate reduce pentru a obține aceeași precizie a rezultatelor. De asemenea, prin apelul la un plan de sondaj mixt, sunt șanse să scadă costurile și timpul de realizare a anchetei.

În opțiunea pentru sondajele mixte, un rol important îl pot avea *natura* populației și *tema* de cercetare. Dacă populația studiată admite o structură particulară sau permite organizarea pe grupe cu o omogenitate ridicată, atunci este preferabil să se ia în considerare aceste elemente, iar eșantionul să păstreze asemenea coordonate. De exemplu, se poate lua în considerare structura pe medii a gospodăriilor, structura teritorială a ariilor urbane (cartiere, zone), tipul locuințelor din orașe (case și blocuri), regiunile de dezvoltare economică, regiunile istorice, zonele geografice și climaterice etc.

Sondajele empirice pot fi obținute mai repede *în timp* și cu un *cost* mai redus, însă principala problemă o reprezintă imposibilitatea de a măsura gradul de precizie al rezultatelor, așa cum se întâmplă în cazul sondajelor aleatoare. Prin combinarea celor două metode, se obține o metodă de sondaj pentru care se pot estima erorile de sondaj, iar rezultatele sunt garantate cu o anumită probabilitate. Chiar dacă există condiții de cost și de timp impuse, prin mixarea celor două metode de sondaj clasice, se obțin rezultate cu o precizie foarte bună, uneori superioară oricăreia dintre cele două metode clasice, iar, în plus, se obține și o măsură a erorilor de sondaj.

Trecerea de la sondajul aleator simplu la cel mixt presupune o analiză a manierei în care se modifică: etapele planului de sondaj, în special extragerea eșantionului din populația de referință; relațiile și proprietățile estimatorilor utilizați pentru estimarea parametrilor; precizia rezultatelor; volumul eșantionului în funcție de o anumită precizie dorită; costul și durata de realizare a sondajului. Aceste aspecte vor fi discutate în următoarele subcapitole pentru sondajul stratificat, sondajul în două sau mai multe trepte și sondajul pe serii.

4.2. Sondajul stratificat

Sondajul aleator simplu presupune extragerea unui eșantion din populația de referință fără a lua în considerare elemente privind structura acelei populații sau gradul său de omogenitate. Nivelul de precizie al rezultatelor obținute cu ajutorul sondajului aleator simplu depinde de volumul eșantionului și de nivelul variației populației. În acest caz, creșterea preciziei se poate realiza fie prin creșterea volumului eșantionului, fie prin reducerea gradului de eterogenitate a populației. Cea de-a doua posibilitate stă la baza metodei de sondaj prezentate în această secțiune.

Cel mai cunoscut și aplicat sondaj mixt este sondajul *stratificat*. Numele provine de la ideea de bază a acestei metode, și anume de a realiza selecția aleatoare după ce populația de referință a fost structurată în grupe sau *straturi* cu ajutorul unei variabile de control puternic corelate cu variabila de interes.

4.2.1. Principiul metodei

Sondajul stratificat are la bază ideea de a combina principiul selecției aleatoare cu cel al eșantionării empirice prin utilizarea unei informații suplimentare despre structura populației. Practic, putem vorbi despre o combinație între sondajul aleator simplu și sondajul pe cote. Informația suplimentară conduce la o structurare a populației de referință pe grupe sau straturi care au o omogenitate mai ridicată decât întregul. De exemplu, în cazul unui sondaj de opinie la nivelul României, există mai multe posibilități de stratificare: pe medii de proveniență (rural și urban); pe regiuni de dezvoltare; pe județe; după mărimea localităților etc.

Rezultatul stratificării depinde de omogenitatea straturilor în raport cu variabila de interes. Principiul stratificării vizează asigurarea unei reprezentativități la nivel de eșantion a unei structuri existente la nivelul populației. Acest lucru presupune obținerea unei omogenități ridicate la nivelul fiecărui strat și realizarea unor diferențe semnificative între straturi. Variabila după care se realizează gruparea trebuie să fie puternic corelată cu tema de cercetare. Structurarea populației pe straturi presupune cunoașterea ponderilor grupelor din total sau a numărului de unități din populație din fiecare strat. În practică, aceste tipuri de sondaje au la bază o structurare a populației pe criterii cum ar fi cele geografice (regiuni sau arii cu trăsături specifice), administrative (unități administrativ teritoriale), economice (regiuni de dezvoltare, sectoare de activitate) etc. Metoda stratificării conduce la atingerea a două obiective: creșterea calității estimațiilor obținute; asigurarea reprezentativității eșantionului la nivelul unor categorii distincte de unități din populație (la nivel de straturi) ce sunt considerate de mare importanță pentru cercetare.

Metoda sondajului stratificat presupune construirea unui eșantion în două etape (vezi tabelul 4.1). Mai întâi se realizează o partiționare sau stratificare a

populației cu ajutorul unei variabile de grupare. Aceste straturi pot avea același număr de unități sau mărimea lor poate să difere. În pasul al doilea se aplică metoda selecției aleatoare simple la nivelul fiecărui strat. Eșantionul final se obține ca urmare a acestor selecții de sub-eșantioane independente, iar condițiile de calitate ale acestuia se pot determina ușor pe baza acestei proprietăți.

Tabelul 4.1. *Principiul selecției stratificate*

Nr.	**Straturi** N_j	**Extragere aleatoare** n_j
1	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	○ ○ ○ ○ ● ○ ○ ● ○ ○ ●
2	○ ○ ○ ○ ○ ○ ○	○ ○ ● ○ ○ ● ○
3	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	○ ○ ● ○ ● ● ○ ○ ○ ○ ○ ● ○
4	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	● ○ ○ ● ● ○ ● ○ ○ ○ ○
...	...	...
H	○ ○ ○ ○ ○	○ ● ○ ●○
Total	N	n

Condiția de independență a extragerii unui număr de unități din fiecare strat permite utilizarea unor metode de selecție diferite pentru fiecare strat în parte. Mai precis, din unele straturi se poate extrage aleator simplu, iar din altele se poate aplica extragerea cu probabilități inegale. Această proprietate arată flexibilitatea sondajelor stratificate, iar modalitățile de extragere de la nivelul fiecărui strat depind de informațiile suplimentare existente și de natura straturilor, așa cum au fost construite cu ajutorul variabilei de grupare. În această lucrare, considerăm aceeași schemă de extragere aleatoare simplă din toate straturile. Mai trebuie subliniat că independența extragerilor este importantă pentru construcția estimatorilor parametrilor, în special pentru varianța acestora (varianța la nivel de eșantion se obține ca sumă a variațelor la nivel de straturi).

Sondajul stratificat asigură ameliorarea calității unui sondaj aleator simplu prin reducerea varianței estimatorilor, adică se realizează o creștere a preciziei rezultatelor. De asemenea, prin creșterea nivelului de omogenitate la nivelul straturilor, se poate reduce volumul eșantionului, comparativ cu sondajul aleator simplu, ceea ce presupune un cost mai redus și un timp de realizare mai mic.

Notăm cu X variabila de stratificare şi cu Y variabila de interes. De asemenea, notăm cu H numărul de straturi, iar cu N_h volumul fiecărui strat, respectând condiția:

$$\sum_{h=1}^{H} N_h = N$$

Pe lângă necesitatea cunoașterii ponderilor $\frac{N_h}{N}$, sondajul stratificat presupune realizarea unei baze de sondaj la nivelul fiecărui strat. Eșantionul se obține prin extragerea aleatoare simplă a câte unui sub-eșantion de unități n_h din fiecare strat, respectând condiția:

$$\sum_{h=1}^{H} n_h = n$$

Deși cunoașterea structurii populației după o variabilă puternic corelată cu tema de cercetare impune un cost în plus, această situație poate fi compensată prin câştigul adus de facilitarea procesului de observare şi prin reducerea volumului eşantionului.

Prin stratificare, varianța totală a variabilei de interes se descompune în două componente: varianța explicată (*between*) sau varianța mediilor condiționate de la nivelul straturilor şi varianța reziduală (*within*) sau media varianțelor condiționate de la nivelul fiecărui strat:

$$V(Y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2 = \sum_{h=1}^{H}\frac{N_h}{N}(\mu_h - \mu)^2 + \sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2$$

unde

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\mu_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi}$$

$$\sigma_h^2 = \frac{1}{N_h}\sum_{i=1}^{N_h}(y_{hi} - \mu_h)^2$$

Ca urmare a principiului stratificării, în analiza proprietăților estimatorilor se ia în calcul doar varianța de la nivelul straturilor, adică varianța intra-strat. Stratificarea este cu atât mai eficientă cu cât straturile sunt mai omogene sau

variantele la nivel de straturi sunt mai mici. De regulă, stratificarea are impact asupra erorilor de estimare, acestea fiind mai mici decât în cazul sondajului aleator simplu, pentru că varianța intra-straturi este mai mică decât varianța totală.

Problemele stratificării

Stabilirea variabilei de stratificare. Stratificarea aduce rezultate cu atât mai semnificative cu cât prin structurarea populației se realizează o varianță între straturi mai mare (preia cât mai mult din varianța totală) și deci o omogenitate mai mare în interiorul straturilor. Problema cea mai dificilă o constituie faptul că variabila de stratificare trebuie cunoscută la nivelul populației de referință: aceasta asigură repartizarea fiecărei unități individuale într-un strat de volum N_h. Cunoașterea informației suplimentare la nivelul populației totale reprezintă o restricție importantă, atât de cost, cât și de posibilitate practică. Din punct de vedere teoretic, este convenabil un număr cât mai mare de straturi, însă, din punct de vedere practic, acesta trebuie limitat la un număr care să răspundă atât condițiilor economice, cât și celor de precizie.

Stabilirea numărului de straturi. În primul rând, numărul de straturi depinde de variabila de stratificare, de variantele sau valorile posibile ale acestei variabile. În practică, însă, numărul straturilor este ales în funcție de mai multe criterii:

- costul - care crește odată cu creșterea numărului de straturi;
- precizia rezultatelor - care scade odată cu creșterea numărului de straturi;
- posibilitățile de eșantionare - nu pot exista straturi cu o singură unitate (nu se poate calcula varianța) sau chiar cu un număr mic de unități.

Repartizarea eșantionului pe straturi. După ce a fost aleasă variabila de stratificare și s-a fixat numărul de straturi, rămâne să se aleagă metoda prin care se compune eșantionul final din cele H sub-eșantioane care se extrag aleator din fiecare strat. Această problemă, numită *alocare*, presupune stabilirea procedeului după care se determină valorile n_h. Asupra acestui subiect ne oprim într-o secțiune ulterioară.

Observație

În cazul sondajului stratificat, numărul de eșantioane posibil de extras din populație este mai mic decât cel din cazul sondajului aleator simplu:

$$K = C_{N_1}^{n_1} \cdot C_{N_2}^{n_2} \cdot \ldots \cdot C_{N_H}^{n_H} \leq C_N^n$$

Ca o concluzie la cele prezentate mai sus, prin stratificare se vizează o serie de avantaje:

- creșterea preciziei rezultatelor prin reducerea erorilor de estimare;
- o mai bună calitate a procesului de culegere a datelor prin utilizarea unor strategii diferite în funcție de natura straturilor;

- reducerea costului unitar prin reducerea costurilor de transport și culegere a datelor;
- obținerea de rezultate (estimații ale parametrilor și ale erorilor) și la nivel de straturi, și la nivelul populației totale;
- utilizarea unor procedee de selecție și a unor metode de culegere a datelor, care pot fi diferite la nivel de strat;
- creșterea nivelului de reprezentativitate a eșantionului;
- reducerea timpului de realizare a anchetei.

4.2.2. Estimarea parametrului medie

Pentru variabila de interes Y, parametrul medie este:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$$

În condițiile stratificării, media se poate determina în funcție de valorile variabilei la nivel de straturi și de mediile la nivel de strat:

$$\mu = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}\frac{1}{N_h} y_{hi}$$

$$\mu = \frac{1}{N}\sum_{h=1}^{H} \mu_h N_h$$

unde

$$\mu_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi}$$

reprezintă media variabilei de interes la nivelul stratului h.

Pentru sondajul stratificat, un estimator nedeplasat pentru μ se obține ca o medie de forma:

$$\hat{\mu}_{st} = \frac{1}{N}\sum_{h=1}^{H} \hat{\mu}_h N_h$$

O valoare a estimatorului $\hat{\mu}_{st}$ la nivelul unui eșantion este o estimație ce se obține cu ajutorul relației:

$$\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{H} \bar{y}_h N_h$$

Se poate observa că estimația mediei se obține ca o medie a mediilor de sondaj pentru cele H sub-eșantioane, ponderate cu ponderile straturilor la nivelul populației.

În cazul unui sondaj aleator simplu, estimatorul mediei este media de selecție:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

cu Y_i variabile aleatoare de selecție independente și identic distribuite cu variabila de interes la nivelul populației totale.

Acest estimator se mai poate construi cu ajutorul estimatorilor construiți pentru fiecare sub-eșantion, potrivit selecției aleatoare simple. Notăm acești estimatori cu $\hat{\mu}_h$, $h = \overline{1,H}$.

Se poate demonstra că estimatorii mediilor la nivel de strat sunt nedeplasați și convergenți. Au loc relațiile:

$$E(\hat{\mu}_h) = \mu_h$$

$$V(\hat{\mu}_h) = \frac{N_h - n_h}{N_h}\frac{\sigma_h^2}{n_h}$$

Pentru sondajului aleator simplu, estimatorul mediei populație este:

$$\hat{\mu} = \frac{1}{n}\sum_{h=1}^{H} \hat{\mu}_h n_h$$

iar estimația, media de eșantion, este:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{h=1}^{H} \bar{y}_h n_h$$

Observație

Se poate observa imediat că estimația punctuală a mediei diferă de media eșantionului obținut prin stratificare:

$$\bar{y} = \frac{1}{n}\sum_{h=1}^{H} \bar{y}_h n_h$$

Rezultă că media de eșantion nu este estimația punctuală a parametrului medie. Evident, ceea ce diferă în cele două expresii sunt ponderile.

Dacă se consideră ponderile la nivel de populație și eșantion egale:

$$\frac{N_h}{N} = \frac{n_h}{n} = f$$

atunci $\hat{\mu}_{st} = \hat{\mu}$, respectiv $\bar{y}_{st} = \bar{y}$.

În acest caz, sondajul stratificat se numește *reprezentativ* sau *proporţional*, iar media de eșantion este estimația punctuală a parametrului medie.

a. Proprietăţile estimatorului mediei

Am arătat că un estimator al medie pentru sondajul stratificat este:

$$\hat{\mu}_{st} = \frac{1}{N}\sum_{h=1}^{H} \hat{\mu}_h N_h$$

Se poate arăta că estimatorul este nedeplasat:

$$E(\hat{\mu}_{st}) = E\left(\frac{1}{N}\sum_{h=1}^{H} \hat{\mu}_h N_h\right) = \frac{1}{N}\sum_{h=1}^{H} E(\hat{\mu}_h) N_h = \frac{1}{N}\sum_{h=1}^{H} \mu_h N_h = \mu$$

Potrivit principiului stratificării, în varianţa estimatorului se iau în calcul doar variantele grupelor sau straturilor. De asemenea, în determinarea erorilor de estimare se ia în considerare doar extragerea aleatoare fără revenire, pentru că prin stratificare se obțin sub-populații cu un volum N_h considerabil redus. Condiția de independență a extragerilor permite determinarea variantei estimatorului mediei după relația:

$$V(\hat{\mu}_{st}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 V(\hat{\mu}_h) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h}\frac{\sigma_h^2}{n_h} = \sum_{h=1}^{H} \frac{N_h^2}{N^2}\frac{N_h - n_h}{N_h}\frac{\sigma_h^2}{n_h}$$

sau

$$V(\hat{\mu}_{st}) = \sum_{h=1}^{H} \frac{N_h^2}{N^2}(1 - f_h)\frac{\sigma_h^2}{n_h}$$

unde

$$V(\hat{\mu}_h) = \frac{N_h - n_h}{N_h}\frac{\sigma_h^2}{n_h}$$

este varianţa estimatorului mediei de strat, iar

$$f_h = \frac{n_h}{N_h}$$

este fracția de sondaj la nivel de strat.

b. Estimarea mediei prin interval de încredere

Pentru calculul intervalelor de încredere, se utilizează un estimator al variației de mai sus:

$$\hat{\sigma}^2_{\bar{\mu}_{st}} = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h} \frac{\hat{\sigma}'^2_h}{n_h}$$

unde $\hat{\sigma}'^2_h$ reprezintă varianța de selecție modificată la nivel de strat sau estimatorul variației la nivelul unui strat.

Intervalul de încredere se obține pe baza proprietăților estimatorului pentru parametrul medie, în principal al convergenței în repartiție. În acest sens, se utilizează o statistică Student:

$$t = \frac{\hat{\mu}_{st}}{\hat{\sigma}_{\bar{\mu}_{st}}}$$

iar intervalul de încredere este:

$$\left(\hat{\mu}_{st} \pm t_{\alpha/2} \hat{\sigma}_{\bar{\mu}_{st}}\right)$$

La nivelul unui eșantion, intervalul de încredere estimat este de forma:

$$\left(\bar{y}_{st} \pm t_{\alpha/2} s_{\bar{\mu}_{st}}\right)$$

unde

$$s^2_{\bar{\mu}_{st}} = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h} \frac{s'^2_h}{n_h}$$

iar s'^2_h sunt estimațiile variațelor la nivel de eșantion corespunzătoare variațelor la nivel de strat:

$$s'^2_h = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$$

Observația 1

Rezultatele obținute pentru parametrul medie sunt similare pentru parametrul proporție.

Considerăm parametrul π și proporțiile la nivel de strat π_h.

Estimatorul nedeplasat al proporției este:

$$\hat{\pi}_{st} = \frac{1}{N} \sum_{h=1}^{H} \hat{\pi}_h N_h$$

iar estimația

$$w_{st} = \frac{1}{N}\sum_{h=1}^{H} w_h N_h$$

unde w_h sunt estimațiile proporțiilor la nivel de strat.

Pentru calculul de precizie se utilizează estimatorul varianței:

$$\hat{\sigma}^2_{\hat{\pi}_{st}} = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h} \frac{\hat{\pi}_h(1-\hat{\pi}_h)}{n_h}$$

iar estimația este:

$$s^2_{\hat{\pi}_{st}} = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h} \frac{w_h(1-w_h)}{n_h}$$

Observația 2

În cazul sondajului stratificat, se poate estima și valoarea globală sau totalul. Dacă notăm cu $\hat{\tau}_{st}$ estimatorul totalului în cazul unui sondaj stratificat, iar cu $\hat{\tau}_h$ un estimator nedeplasat al valorii globale la nivel de strat, rezultă:

$$\hat{\tau}_h = \hat{\mu}_h N_h$$

$$V(\hat{\tau}_h) = N_h^2(1-f_h)\frac{\sigma_h^2}{n_h}$$

$$\hat{\tau}_{st} = \sum_{h=1}^{H} \hat{\tau}_h = \sum_{h=1}^{H} \hat{\mu}_h N_h = N \cdot \hat{\mu}_{st}$$

$$V(\hat{\tau}_{st}) = \sum_{h=1}^{H} N_h^2(1-f_h)\frac{\sigma_h^2}{n_h}$$

Observația 3

Dacă se consideră sondajul proporțional, adică $\frac{N_h}{N} = \frac{n_h}{n} = f$, atunci estimatorul mediei este de forma:

$$\hat{\mu}_{st} = \frac{1}{n}\sum_{h=1}^{H} \hat{\mu}_h n_h$$

iar estimația este:

$$\bar{y}_{st} = \bar{y} = \frac{1}{n}\sum_{h=1}^{H} \bar{y}_h n_h$$

Acest estimator este nedeplasat și are varianța:

$$V(\hat{\mu}_{st}) = \frac{1-f}{n}\sum_{h=1}^{H} \frac{N_h}{N}\sigma_h^2 = \frac{1-f}{n}\sum_{h=1}^{H} \frac{n_h}{n}\sigma_h^2$$

c. Câștigul de precizie al stratificării

Se poate demonstra că varianța estimatorului obținut prin stratificare este mai mică decât a estimatorului obținut prin selecție aleatoare simplă. Această diferență se numește *câștig de precizie* datorat stratificării.

Fie $\hat{\theta}$ estimatorul obţinut pentru parametrul θ în cazul sondajului aleator simplu şi $\hat{\theta}_{st}$ estimatorul obţinut prin metoda stratificării. Atunci, câştigul de precizie absolut (C_a) şi relativ (C_r) se pot scrie:

$$C_a = V(\hat{\theta}) - V(\hat{\theta}_{st})$$

$$C_r = \frac{V(\hat{\theta}) - V(\hat{\theta}_{st})}{V(\hat{\theta})}$$

Criteriul după care se va alege variabila de stratificare este următorul: dintre toate variabilele de stratificare posibile, se va alege variabila pentru care câştigul de precizie este cel mai mare.

Pentru exemplificarea câştigului de precizie, considerăm parametrul medie. Estimatorul clasic, obţinut într-un sondaj aleator simplu, este:

$$\hat{\mu} = \frac{1}{n}\sum_{h=1}^{H} \hat{\mu}_h n_h$$

Acest estimator este nedeplasat şi are varianţa:

$$V(\hat{\mu}) = \frac{N-n}{N}\frac{\sigma^2}{n}$$

sau

$$V(\hat{\mu}) = (1-f)\frac{\sigma^2}{n}$$

Conform principiului stratificării, varianţa populaţiei se descompune după relaţia:

$$\sigma^2 = \sum_{h=1}^{H} \frac{N_h}{N}(\mu_h - \mu)^2 + \sum_{h=1}^{H} \frac{N_h}{N}\sigma_h^2$$

Considerând cazul sondajul stratificat proporţional ($\frac{n_h}{N_h} = \frac{n}{N} = f$), putem scrie varianţele celor doi estimatori ai mediei:

$$V(\hat{\mu}) = \frac{1-f}{n}\left[\sum_{h=1}^{H} \frac{N_h}{N}(\mu_h - \mu)^2 + \sum_{h=1}^{H} \frac{N_h}{N}\sigma_h^2\right]$$

$$V(\hat{\mu}_{st}) = \sum_{h=1}^{H} \frac{N_h^2}{N^2}\frac{N_h - n_h}{N_h}\frac{\sigma_h^2}{n_h} = \frac{1-f}{n}\sum_{h=1}^{H} \frac{N_h}{N}\sigma_h^2$$

În aceste condiţii, câştigul de precizie prin stratificare este:

$$C_a = V(\hat{\mu}) - V(\hat{\mu}_{st}) = \frac{1-f}{n}\sum_{h=1}^{H} \frac{N_h}{N}(\mu_h - \mu)^2$$

$$C_r = \frac{V(\hat{\mu}) - V(\hat{\mu}_{st})}{V(\hat{\mu})} = \frac{\frac{1-f}{n}\sum_{h=1}^{H}\frac{N_h}{N}(\mu_h - \mu)^2}{\frac{1-f}{n}\sigma^2} = \frac{\sum_{h=1}^{H}\frac{N_h}{N}(\mu_h - \mu)^2}{\sigma^2}$$

Observaţia 1

Câştigul absolut de precizie este reprezentat de varianţa inter-straturi. Aceasta poate prelua o parte importantă din varianţa totală dacă se realizează o bună stratificare.

Observaţia 2

Câştigul relativ de precizie al stratificării este echivalent cu raportul de corelaţie dintre variabila de interes şi variabila de stratificare. Într-adevăr, la numărător este varianţa explicată sau varianţa dintre grupe, iar la numitor este varianţa totală:

$$C_r = \eta_{YX}^2 = \frac{V(\mu_h)}{V(Y)}$$

În concluzie, precizia obţinută prin stratificare este cu atât mai mare cu cât raportul de determinaţie dintre variabila de interes şi cea de stratificare este mai mare, adică cu cât legătura dintre cele două variabile este mai puternică.

Deoarece $0 \leq \eta_{YX}^2 \leq 1$, rezultă:

- un câştig zero sau nici o diferenţă între stratificare şi sondajul aleator simplu când între cele două variabile nu există nici o legătură;
- un câştig maxim când legătura dintre cele două variabile este funcţională;
- un câştig relativ de precizie prin stratificare dacă $0 < \eta_{YX}^2 < 1$.

Ca o măsură a eficienţei relative a două metode de sondaj (determinată pe baza preciziei), se poate determina un indicator definit prin inversul raportului varianţelor estimatorilor obţinuţi prin cele două metode pentru un anumit parametru (Yates, 1953, p. 265).

De exemplu, eficienţa sau precizia relativă a unui sondaj stratificat faţă de un sondaj aleator este dată de raportul:

$$\frac{V(\hat{\mu})}{V(\hat{\mu}_{st})} = \frac{\frac{1-f}{n}\left[\sum_{h=1}^{H}\frac{N_h}{N}(\mu_h-\mu)^2 + \sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2\right]}{\frac{1-f}{n}\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2}$$

$$\frac{V(\hat{\mu})}{V(\hat{\mu}_{st})} = \frac{\sum_{h=1}^{H}\frac{N_h}{N}(\mu_h-\mu)^2 + \sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2}{\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2}$$

$$\frac{V(\hat{\mu})}{V(\hat{\mu}_{st})} = 1 + \frac{\sum_{h=1}^{H}\frac{N_h}{N}(\mu_h-\mu)^2}{\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2}$$

Din relaţia de mai sus se poate observa că, în general, prin stratificare se obţine un câştig de precizie, cu excepţia cazului în care raportul din ultima relaţie ar fi egal cu zero, ceea ce înseamnă că varianţa dintre straturi este egală cu zero, adică stratificarea nu produce diferenţe între unităţile populaţiei.

Exemplu

Pentru o companie de 10000 de angajaţi, se realizează un studiu privind disponibilitatea acestora de a lucra un număr de ore suplimentare pe săptămână. Se consideră o grupare a angajaţilor firmei după criteriul educaţie, după cum urmează: 10% studii medii, 65% studii superioare de licenţă, 25% studii postuniversitare de master. La nivelul eşantionului extras de 1000 de angajaţi, se cunosc următoarele rezultate privind variabila timp de lucru suplimentar (ore pe săptămână):

	Mediu	**Licenţă**	**Master**
n_h	150	500	350
$\bar{y}_h$	3	2	1
$s_h'^2$	1,44	0,49	0,09

Se cere:

- să se estimeze punctual și prin interval de încredere timpul mediu de lucru suplimentar la nivelul întregii populații, cu o încredere de 0,95;
- să se testeze dacă media populației totale diferă semnificativ de valoarea dorită de angajator de 2 ore pe săptămână ca timp suplimentar de lucru pe angajat.

Conform datelor de mai sus, estimația mediei la nivel de eșantion se calculează astfel:

$$\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{H} \bar{y}_h N_h$$

unde N_h sunt obținute pe baza ponderilor din enunț:

$N_1 = 1000$; $N_2 = 6500$; $N_3 = 2500$.

Rezultă:

$$\bar{y}_{st} = \frac{1}{10000}(3 \cdot 1000 + 2 \cdot 6500 + 1 \cdot 2500) = 1{,}85 \text{ ore}$$

Pentru intervalul de încredere, se utilizează relația:

$$\left(\bar{y}_{st} \pm t_{\alpha/2} s_{\hat{\mu}_{st}}\right)$$

unde

$$s^2_{\hat{\mu}_{st}} = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h} \frac{s_h'^2}{n_h}$$

Din tabelul Student se citește valoarea $t_{\alpha/2} = 1{,}96$, pentru că eșantionul este de volum mare.

Din calcule rezultă $s^2_{\hat{\mu}_{st}} = 0{,}00047$, iar $s_{\hat{\mu}_{st}} = 0{,}021$.

Intervalul de încredere estimat, pentru o probabilitate de 0,95 este:

$$\left(\bar{y}_{st} \pm t_{\alpha} s_{\hat{\mu}_{st}}\right) \Rightarrow (1{,}85 \pm 1{,}96 \cdot 0{,}021) \Rightarrow (1{,}85 \pm 0{,}041) \text{ ore}$$

Pentru a doua cerință, avem cazul unui test Student de egalitate a mediei populației cu o constantă. Ipotezele statistice sunt următoarele:

$$H_0: \mu = 2$$

$$H_1: \mu \neq 2$$

Valoarea calculată a testului se obține astfel:

$$t_{calc} = \frac{\bar{y}_{st} - 2}{s_{\hat{\mu}_{st}}} = \frac{1{,}85 - 2}{0{,}021} = -7{,}14$$

Pentru o încredere de 0,95, se observă că $|t_{calc}| > 1{,}96$, adică se respinge ipoteza nulă.

4.2.3. Construirea eșantionului

În cazul sondajului stratificat, construirea eșantionului presupune câteva etape importante: stabilirea volumului eșantionului; repartizarea acestui efectiv la nivelul straturilor, etapă numită *alocare*; extragerea efectivelor n_h de la nivelul fiecărui strat, conform schemei de extragere stabilite.

a. Calculul volumului eșantionului

Pentru determinarea volumului eșantionului, se utilizează aceleași situații și relații de calcul prezentate pentru cazul sondajului aleator simplu. Singura diferență este aceea că în calcul se ia în considerare doar o componentă a variației totale, și anume varianța intra-straturi notată:

$$\bar{\sigma}^2 = \sum_{h=1}^{H} \frac{N_h}{N} \sigma_h^2$$

De exemplu, dacă variabila de interes este una numerică și se fixează o limită pentru eroarea maxim admisibilă, volumul eșantionului se poate calcula după relația:

$$n = \frac{z^2 \bar{\sigma}^2}{\Delta_{\mu}^2 + \frac{z^2 \bar{\sigma}^2}{N}}$$

iar Δ_{μ} este o valoare stabilită a erorii maxim admisibile pentru parametrul medie.

Deoarece $\bar{\sigma}^2 < \sigma^2$, volumul eșantionului pentru cazul sondajului stratificat este mai mic decât cel calculat pentru cazul sondajului aleator simplu.

Exemplu

Considerăm cazul unui sondaj care studiază consumul lunar de ciocolată pentru o populație de 10000 de persoane. Populația este stratificată pe grupe de vârstă, iar dintr-un studiu anterior se cunoaște varianța intra-strat, care este egală cu 50000. Se cere să se estimeze volumul eșantionului care trebuie extras aleator fără revenire, dacă se dorește să se estimeze consumul mediu de ciocolată cu o eroare standard care nu depășește 10 grame pe o persoană, cu o probabilitate de 0,95.

Se cunosc relațiile:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

$$n_0 = \frac{z^2 \bar{\sigma}^2}{\Delta_\mu^2} = \frac{\bar{\sigma}^2}{SE^2}$$

Înlocuind în relațiile de mai sus, se obține:

$n_0 = 500$ de persoane, iar $n = 476$ de persoane.

b. Alocarea volumului eșantionului

Din punct de vedere teoretic și practic, există cel puțin trei posibilități sau tipuri de alocare: simplă, proporțională, optimală.

Alocarea simplă

Metoda presupune extragerea aleatoare simplă din fiecare strat a unui număr egal de unități care să compună eșantionul. Rezultă că:

$$n_1 = n_2 = \cdots = n_H = \frac{n}{H}$$

În condițiile alocării simple, analiza datelor obținute pe baza unei asemenea metode ar putea permite testarea existenței unor diferențe ce există între straturi, dar asumând ipoteza egalității variațelor la nivel de strat. Dacă o asemenea ipoteză nu este valabilă, pentru verificarea diferențelor dintre straturi pentru parametrul medie, de exemplu, este necesar să se utilizeze o altă metodă de alocare.

Alocarea proporțională

Această metodă pornește de la ideea de a extrage din fiecare strat un număr de unități proporțional cu mărimea stratului:

$$n_h = n \cdot \frac{N_h}{N}$$

În aceste condiții, sondajul se numește *stratificat reprezentativ* sau *proporțional*, deoarece se utilizează fracția de sondaj (f) la nivelul fiecărui strat:

$$\frac{n_h}{n} = \frac{N_h}{N}$$

sau

$$\frac{n_h}{N_h} = \frac{n}{N} = f$$

Pentru estimarea punctuală și prin interval de încredere a medie se utilizează relațiile:

$$\bar{y} = \frac{1}{n}\sum_{h=1}^{H} \bar{y}_h n_h$$

$$\left(\bar{y} \pm t_{\alpha/2} s_{\hat{\mu}}\right)$$

$$s_{\hat{\mu}}^2 = \frac{1-f}{n}\sum_{h=1}^{H} \frac{n_h}{n} s_h'^2$$

În practică, acest tip de alocare este frecvent utilizat datorită simplității aplicării metodei și a facilității calculelor privitoare la estimații și erori, chiar dacă rezultatele nu respectă neapărat condiția de optim în privința preciziei.

Alocarea optimă (Neyman)

În condițiile stratificării, știind că eșantionul este compus din mai multe sub-eșantioane extrase din straturi relativ omogene, se pune problema repartizării volumului eșantionului pe cele H straturi astfel încât să se obțină cea mai bună precizie a estimatorului mediei variabilei de interes. Optimizarea presupune determinarea valorilor n_h care minimizează varianța estimatorului parametrului care se dorește a fi estimat prin sondaj.

Cu alte cuvinte, avem o problemă de optim cu legături de forma:

$$(P)\begin{cases} \min\limits_{n_h} V(\hat{\mu}_{st}) \\ \sum\limits_{h=1}^{H} n_h = n \end{cases}$$

Conform metodei multiplicatorului Lagrange, rezolvarea problemei presupune anularea derivatelor parțiale de ordinul întâi aplicate relației:

$$L(n_1, n_2, \ldots, n_h) = V(\hat{\mu}_{st}) + \lambda\left(\sum_{h=1}^{H} n_h - 1\right)$$

unde

$$V(\hat{\mu}_{st}) = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h} \frac{\sigma_h^2}{n_h}$$

Deci,

$$\frac{\partial L}{\partial n_h} = -\frac{N_h^2}{N^2}\frac{\sigma_h^2}{n_h^2} + \lambda$$

Pentru

$$\frac{\partial L}{\partial n_h} = 0 \Rightarrow -\frac{N_h^2}{N^2}\frac{\sigma_h^2}{n_h^2} + \lambda = 0$$

adică

$$n_h^2 = \frac{1}{\lambda N^2} N_h^2 \sigma_h^2$$

sau

$$n_h = \frac{1}{\sqrt{\lambda} N} N_h \sigma_h$$

Din ultima relaţie se observă că minimul varianţei estimatorului se realizează în condiţiile în care volumul fiecărui sub-eşantion este proporţional cu abaterea medie pătratică din stratul din care a fost extras.

Ştiind că $\sum_{h=1}^{H} n_h = n$, din ultima relaţie se poate observa că:

$$\lambda = \left(\frac{\sum_{h=1}^{H} N_h \sigma_h}{n \cdot N} \right)^2$$

iar prin înlocuire se obţine relaţia repartiţiei optime a eşantionului pe cele H straturi sau alocarea optimală Neyman:

$$n_h = n \frac{N_h \sigma_h}{\sum_{h=1}^{H} N_h \sigma_h}$$

Relaţia de mai sus ne oferă o relaţie de proporţionalitate:

$$\frac{n_h}{N_h \sigma_h} = \frac{n}{\sum_{h=1}^{H} N_h \sigma_h} = const.$$

În mod analog, dacă se consideră costul anchetei fixat dinainte, se poate realiza o alocare în funcţie de costul unitar de sondaj la nivelul fiecărui strat (c_h):

$$\frac{n_h}{\frac{N_h \sigma_h}{\sqrt{c_h}}} = \frac{C}{\sum_{h=1}^{H} N_h \sigma_h \sqrt{c_h}} = const.$$

unde

$$C = \sum_{h=1}^{H} c_h n_h$$

reprezintă costul total al anchetei şi este fixat.

Rezultă:

$$n_h = n \frac{N_h \sigma_h / \sqrt{c_h}}{\sum_{h=1}^{H} (N_h \sigma_h / \sqrt{c_h})}$$

În cazul alocării optimale, se poate demonstra că estimaţia varianţei estimatorului mediei este de forma:

$$s_{\bar{\mu}}^2 = \frac{1}{n} \left(\sum_{h=1}^{H} \frac{N_h}{N} s_h' \right)^2 - \frac{1}{N} \sum_{h=1}^{H} \frac{N_h}{N} s_h'^2$$

iar aceasta este mai mică decât în cazul alocării proporţionale.

Observaţii

1. Din fiecare strat se va extrage un sub-eşantion cu atât mai mare cu cât:
- volumul stratului este mai mare;
- varianţa stratului este mai mare;
- costul unitar de anchetă din fiecare strat este mai mic.

Limita acestor alocări optimale este determinată de cunoaşterea varianţelor din fiecare strat, precum şi a costurilor unitare din fiecare strat.

2. Dacă eşantionul obţinut este optim pentru o variabilă de interes Y, nimic nu garantează că este la fel şi pentru o altă variabilă. În aceste condiţii, este preferabil să se utilizeze sondajul stratificat proporţional.

3. Dacă se înregistrează acelaşi cost unitar în fiecare strat, alocarea optimală se realizează doar în funcţie de varianţa din fiecare strat.

4. Câştigul de precizie adus de alocarea optimală este important doar dacă varianţele de la nivelul straturilor diferă semnificativ între ele.

5. Alocarea optimală depinde de cunoaşterea varianţei fiecărui strat după variabila de interes. Această constrângere poate ridica probleme deosebite în practică, deoarece aceste mărimi de obicei nu sunt cunoscute. Varianţele pentru fiecare strat pot fi estimate fie pe baza unor informaţii din anchete anterioare, fie prin realizarea unei pre-anchete.

Exemplu

Pentru o companie de 10000 de angajaţi, se realizează un studiu privind disponibilitatea acestora de a lucra un număr de ore suplimentare pe săptămână. Se

consideră o grupare a angajaților firmei după criteriul educație, după cum urmează: 10% studii medii, 65% studii superioare de licență, 25% studii postuniversitare de master. La nivelul eșantionului extras de 1000 de angajați, se cunosc următoarele rezultate privind variabila timp de lucru suplimentar (ore pe săptămână):

	Mediu	**Licență**	**Master**
n_h	150	500	350
$\bar{y}_h$	3	2	1
$s_h'^2$	1,44	0,49	0,09

Se cere:

- să se estimeze punctual și prin interval de încredere parametrul medie în condițiile unui sondaj stratificat proporțional;
- să se testeze dacă media populației diferă semnificativ de 2 ore pe săptămână ca timp suplimentar de lucru.

În acest caz, estimația mediei este identică cu media obținută la nivelul eșantionului:

$$\bar{y} = \frac{1}{n}\sum_{h=1}^{H} \bar{y}_h n_h$$

De asemenea, în cazul alocării proporționale, se folosește estimatorul:

$$\hat{\sigma}_{\bar{y}}^2 = \frac{1-f}{n}\sum_{h=1}^{H} \frac{N_h}{N}\hat{\sigma}_h'^2$$

iar estimația este:

$$s_{\bar{y}}^2 = \frac{1-f}{n}\sum_{h=1}^{H} \frac{n_h}{n}s_h'^2$$

Din calcule rezultă:

$$\bar{y} = \frac{1}{n}\sum_{h=1}^{H} \bar{y}_h n_h = \frac{1}{1000}(3 \cdot 150 + 2 \cdot 500 + 1 \cdot 350) = 1{,}8 \text{ ore}$$

$$s_{\hat{\mu}} = 0{,}021 \text{ ore}$$

Intervalul de încredere estimat, pentru o probabilitate de 0,95 este:

$$\left(\bar{y} \pm t_{\frac{\alpha}{2}} \cdot s_{\hat{\mu}}\right)$$

Adică rezultă intervalul estimat:

$$(1{,}80 \pm 1{,}96 \cdot 0{,}021) \Rightarrow (1{,}80 \pm 0{,}041) \text{ ore}$$

Pentru ultima cerință, avem cazul unui test Student de egalitate a mediei populației cu o constantă. Ipotezele statistice sunt următoarele:

$$H_0: \mu = 2$$

$$H_1: \mu \neq 2$$

Valoarea calculată a testului se obține astfel:

$$t_{calc} = \frac{\bar{y} - 2}{s_{\hat{\mu}}} = \frac{1{,}8 - 2}{0{,}021} = -9{,}5$$

Pentru o încredere de 0,95, se observă că $|t_{calc}| > 1{,}96$, adică se respinge ipoteza nulă.

4.3. Sondajul în două sau mai multe trepte

Sondajul stratificat presupune două etape metodologice, însă doar în una dintre acestea se apelează la selecția aleatoare. Există situații când principiul aleator de selecție se poate aplica la mai multe niveluri în cadrul unui plan de sondaj mixt. Asemenea situații sunt întâlnite în cazul sondajelor realizate în două sau mai multe trepte, numărul acestora fiind dat de frecvența cu care se apelează la principiul selecției aleatoare. În această secțiune, prezentăm câteva aspecte legate de construirea unui plan de sondaj în două trepte ce utilizează principiul selecției aleatoare simple în ambele etape ale selecției.

4.3.1. Fundamentarea metodei

Ideea construirii unui eşantion printr-o selecție în două etape este strâns legată de natura populațiilor studiate. De exemplu, pentru analiza fenomenelor economice și sociale se ia în considerare frecvent organizarea administrativ teritorială a unei țări sau regiuni geografice. Pentru construirea eșantionului se ia în calcul împărțirea pe județe, orașe, comune, iar în interiorul fiecărei unități administrative se pot identifica alte structuri care sunt relevante pentru asigurarea reprezentativității: cartiere, străzi, clădiri etc.

Prin organizarea unui sondaj în două sau mai multe trepte, se au în vedere următoarele limite ale sondajului aleator simplu (Ardilly, 1994, p. 94):

- costurile de culegere a datelor, mai ales în cazurile în care eşantionul este foarte dispersat în teritoriu;
- construirea bazei de sondaj, care trebuie să cuprindă toate unităţile populaţiei şi să conţină elementele de identificare a fiecărei unităţi în parte.

Principiul de bază al sondajului în două trepte este structurarea populaţiei în grupe mari de unităţi numite *unităţi primare* (UP). Fiecare unitate a populaţiei este identificată prin apartenenţa la una dintre aceste UP. În consecinţă, baza de sondaj

este constituită la două niveluri: există o bază de sondaj a *UP*, iar fiecare *UP* are o bază de sondaj pentru unitățile individuale care o compun, numite și *unități secundare* (*US*). De exemplu, într-o anchetă privind activitatea agricolă a gospodăriilor, *UP* pot fi regiuni sau unități administrativ teritoriale, iar ca *US* se pot considera gospodăriile.

Extragerea eșantionului se realizează în două trepte (a se vedea tabelul 4.2):

- mai întâi, se extrag aleator, după o procedură cunoscută, *UP* și se obține un eșantion de *UP* (de exemplu, se extrag cele trei *UP* marcate în tabel). Este important de subliniat că obținerea eșantionului presupune construirea unor baze de sondaj doar pentru *UP* extrase deja;
- din fiecare *UP* se extrag aleator *US*, după o procedură care nu depinde de prima extragere. *US* astfel extrase compun eșantionul final (de exemplu, se extrag cele 6 *US* marcate în tabel). Extragerile din fiecare etapă pot fi cu probabilități egale sau inegale. Decizia se ia în funcție de informația disponibilă cu privire la structura populației pe *UP*.

Tabelul 4.2. *Principiul selecției în două trepte*

UP	**Etapa 1 - extragere *UP***	**Etapa 2 - extragere *US***
○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
○ ○ ○ ○ ○	○○ ○ ○ ○	○● ○ ● ○
○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
○ ○ ○ ○ ○ ○ ○ ○ ○ ○	○ ○ ○ ○ ○ ○ ○ ○ ○ ○	○ ○ ● ○ ○ ○ ● ○ ○ ●
...	...	...
○ ○ ○ ○ ○	○ ○ ○ ○ ○	○ ● ○ ○ ○

Prin acest procedeu de selecție se elimină inconvenientele prezentate mai sus: eșantionul nu mai este atât de dispersat și nu mai este nevoie de o bază de sondaj pentru întreaga populație. Metoda poate fi aplicată și în cazul în care bazele de sondaj cu *US* lipsesc parțial sau total. După extragerea *UP*, se pot realiza cercetări exhaustive în interiorul acestor unități (cazul sondajului pe serii) sau se pot constitui baze de sondaj doar după ce sunt cunoscute *UP* extrase. În aceste condiții, costurile anchetei se pot reduce considerabil.

Avantajul metodei sondajului în două trepte se poate evidenţia în cazul anchetelor organizate în gospodării, sondaje care, de regulă, au caracter periodic. În aceste condiţii, se extrag *UP*, care pot rămâne fixe, iar unităţile secundare pot fi extrase din nou la fiecare anchetă. De asemenea, se poate fixa şi echipa de operatori, care va administra chestionarul în zone restrânse. O importantă reducere a costurilor, precum şi o precizie a rezultatelor se asigură şi prin realizarea unor baze de sondaj în cadrul acestor grupe primare selectate. Bazele de sondaj astfel construite sunt mai facil de administrat şi de actualizat, iar erorile de eşantionare şi de observare pot fi identificate şi reduse mult mai uşor.

Metoda prezintă şi un dezavantaj, numit *efect de sondaj* (Ardilly, 1994, p. 96). Este vorba despre efectul redundanţei informaţiei care apare din cauză că unităţile din eşantion provin doar din anumite unităţi primare extrase în prima treaptă a sondajului. În final, rezultatele sondajului în două trepte sunt mai puţin precise decât rezultatele unui sondaj aleator, însă costul şi timpul de realizare sunt mult reduse.

Condiţiile de eficienţă ale sondajului în două trepte impun ca *UP* să aibă următoarele caracteristici (Desabie, 1966, p. 263):

- volumul lor să nu fie prea mare, deci să existe un număr suficient de mare de asemenea unităţi, astfel încât să fie posibilă o extragere aleatoare a acestora în prima treaptă;
- o relativă omogenitate în raport cu variabilele de interes. Dacă nu există această omogenitate, împărţirea populaţiei în grupe primare nu este eficientă, iar rezultatele nu sunt precise;
- diferenţele de volum între unităţi să nu fie mari.

În concluzie, calitatea rezultatelor depinde de numărul de *UP* extrase în cadrul eşantionului, de numărul de *US* extrase din fiecare *UP* şi de procedeul de extragere din fiecare treaptă.

4.3.2. Proprietăţile estimatorilor. Precizia rezultatelor

Pentru evaluarea preciziei rezultatelor sondajului în două trepte, este important să se ia în considerare principiul descompunerii varianţei totale a variabilei de interes (Y) în două componente determinate de cele două etape ale procesului de selecţie (caz întâlnit şi la sondajul stratificat). Diferenţa faţă de sondajul stratificat este că, în acest caz, pentru calculul preciziei, contează ambele componente ale varianţei, atât din interiorul *UP*, cât şi între acestea, pentru că selecţia aleatoare se realizează la ambele niveluri.

Notăm cu M numărul de *UP* şi cu N_j numărul de unităţi individuale la nivelul fiecărei *UP* şi cu N numărul total de unităţi la nivelul populaţiei, iar cu Y variabila de interes şi cu X variabila după care se realizează structurarea pe *UP*.

La nivelul unei *UP*, media variabilei de interes este:

$$\mu_j = \frac{1}{N_j}\sum_{i=1}^{N_j} y_{ij}$$

$$\sum_{j=1}^{M} N_j = N = M \cdot \bar{N}$$

$$\bar{N} = \frac{N}{M}$$

unde $\bar{N}$ este numărul mediu de unități pe UP.

Similar, se poate construi totalul la nivelul UP:

$$\tau_j = \sum_{i=1}^{N_j} y_{ij} = \mu_j \cdot N_j$$

Dacă se cunoaște totalul, se obține media US la nivelul unei anumite US_j:

$$\mu_j = \frac{\tau_j}{N_j}$$

Totalul mediu la nivelul unei UP este dat de relația:

$$\bar{\tau} = \frac{\tau}{M} = \frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N_j} y_{ij}$$

Valoarea totală la nivelul populației este:

$$\tau = \sum_{i=1}^{N} y_i = \sum_{j=1}^{M} \tau_j = \sum_{j=1}^{M}\sum_{i=1}^{N_j} y_{ij} = N \cdot \mu$$

De asemenea, media la nivelul populației totale pe US este dată de relația:

$$\mu = \frac{1}{N}\sum_{j=1}^{M} \mu_j N_j = \frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N_j} y_{ij} = \frac{\bar{Y}}{\bar{N}}$$

iar varianța totală a variabilei de interes este:

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N_j} (y_{ij} - \mu)^2$$

Pentru procesul de estimare se iau în considerare următoarele variante ce se obțin pentru structurarea populației pe UP și US:

- componenta *within* sau varianța valorilor la nivel de US din interiorul tuturor UP (σ_w^2):

$$\sigma_w^2 = \frac{1}{M}\sum_{j=1}^{M}\sigma_j^2$$

unde

$$\sigma_j^2 = \frac{1}{N_j - 1}\sum_{i=1}^{N_j}\left(y_{ij} - \mu_j\right)^2$$

- componenta *between* se poate scrie în două moduri, în funcție de totalurile și de mediile la nivel de UP:

$$\sigma_b^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(\tau_j - \bar{\tau}\right)^2$$

$$\sigma_b'^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(\mu_j - \mu\right)^2$$

La nivelul unui eșantion de volum n, notăm cu m numărul de UP selectate, iar cu n_j numărul de US selectate din UP_j. Rezultă:

$$n = \sum_{j=1}^{m} n_j$$

Media US selectate de la nivelul unei UP este:

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$$

iar variantele estimate sunt:

$$s_j'^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(y_{ij} - \bar{y}_j\right)^2$$

$$s_w^2 = \frac{1}{m}\sum_{j=1}^{m} s_j'^2$$

$$s_b^2 = \frac{1}{m-1}\sum_{j=1}^{m}(T_j - \bar{T})^2$$

$$s_b'^2 = \frac{1}{m-1}\sum_{j=1}^{m}(\bar{y}_j - \bar{y})^2$$

Totalul la nivelul eșantionului extras dintr-o UP este de forma:

$$T_j = N_j \cdot \bar{y}_j = \frac{N_j}{n_j}\sum_{i=1}^{n_j} y_{ij}$$

Media eșantionului se obține ca o medie a mediilor de la nivelul sub-eșantioanelor extrase din fiecare UP:

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{m}\bar{y}_j n_j = \frac{1}{n}\sum_{i=1}^{n} y_i$$

a. Estimatori ai totalului și mediei

În principiu, construirea unui estimator pentru un anumit parametru θ trebuie să țină cont de cele două etape ale selecției aleatoare. Așadar, media estimatorului se obține astfel:

$$E(\hat{\theta}) = E_1\left[E_2(\hat{\theta})\right]$$

unde E_1 și E_2 sunt operatorii medie pentru fiecare etapă a selecției, adică pentru toate UP ce pot fi extrase în prima treaptă și pentru toate US ce pot fi selecționate în a doua treaptă, în condițiile în care au fost selecționate deja un număr de UP.

Similar, varianța presupune proprietatea:

$$V(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = E_1E_2(\hat{\theta} - \theta)^2 = V_1\left[E_2(\hat{\theta})\right] + E_1\left[V_2(\hat{\theta})\right]$$

Un estimator nedeplasat pentru totalul la nivel de UP

$$\tau_j = \sum_{i=1}^{N_j} y_{ij} = \mu_j \cdot N_j$$

este construit pe baza estimatorului mediei la nivel de UP:

$$\hat{\tau}_j = \hat{\mu}_j \cdot N_j$$

$$\hat{\tau}_j = \frac{N_j}{n_j}\sum_{i=1}^{n_j} Y_{ij}$$

unde Y_{ij} sunt variabile aleatoare de selecție construite prin extragerea aleatoare a US din UP_j. Estimatorul totalului la nivelul populației este:

$$\hat{\tau} = \frac{M}{m}\sum_{j=1}^{m}\hat{\tau}_j = \frac{M}{m}\sum_{j=1}^{m}\frac{N_j}{n_j}\sum_{i=1}^{n_j}Y_{ij}$$

iar estimația sa are relația:

$$T = \frac{M}{m}\sum_{j=1}^{m}T_j = \frac{M}{m}\sum_{j=1}^{m}\frac{N_j}{n_j}\sum_{i=1}^{n_j}y_{ij}$$

sau

$$T = N \cdot \bar{y}$$

Un estimator nedeplasat pentru media populației este:

$$\hat{\mu} = \frac{1}{N}\hat{\tau} = \frac{M}{Nm}\sum_{j=1}^{m}\hat{\tau}_j = \frac{M}{Nm}\sum_{j=1}^{m}\hat{\mu}_j \cdot N_j = \frac{M}{N}\frac{1}{m}\sum_{j=1}^{m}\frac{N_j}{n_j}\sum_{i=1}^{n_j}Y_{ij}$$

Aplicând operatorul medie relației de mai sus, se obține ușor proprietatea de nedeplasare:

$$E(\hat{\mu}) = \mu$$

Se poate demonstra că varianţa estimatorului totalului $\hat{\tau}$ este:

$$V(\hat{\tau}) = M^2(1-f_1)\cdot\frac{\sigma_b^2}{m}\cdot + \frac{M}{m}\sum_{j=1}^{m}N_j^2(1-f_{2j})\frac{\sigma_j^2}{n_j}$$

sau

$$V(\hat{\tau}) = M^2\left(1-\frac{m}{M}\right)\cdot\frac{1}{m}\cdot\frac{1}{M-1}\sum_{j=1}^{M}(\tau_j-\bar{\tau})^2 + \frac{M}{m}\sum_{j=1}^{m}N_j^2\left(1-\frac{n_j}{N_j}\right)\frac{\sigma_j^2}{n_j}$$

În relația de mai sus apar două ponderi sau două fracții de sondaj specifice celor două etape ale eșantionării:

$$f_1 = \frac{m}{M}$$

cu privire la UP și prima treaptă de selecție;

$$f_{2j} = \frac{n_j}{N_j}$$

cu privire la *US*.

O estimație a varianței estimatorului pentru total este de forma:

$$s_{\hat{t}}^2 = M^2\left(1-\frac{m}{M}\right)\cdot\frac{1}{m}\cdot\frac{1}{m-1}\sum_{j=1}^{M}\left(T_j-\frac{T}{M}\right)^2+\frac{M}{m}\sum_{j=1}^{m}N_j^2\left(1-\frac{n_j}{N_j}\right)\frac{s_j'^2}{n_j}$$

unde

$$s_j'^2=\frac{1}{n_j-1}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2$$

Un estimator nedeplasat al varianței mediei de selecție se obține prin relația:

$$V(\hat{\mu})=\frac{1}{N^2}V(\hat{\tau})$$

Analog,

$$s_{\hat{\mu}}^2=\frac{1}{N^2}s_{\hat{t}}^2$$

Observații

1. Dacă fracția de sondaj în extragerea din treapta a doua este constantă, adică $\frac{n_j}{N_j}=f_2$, iar *UP* au același volum $N_j=\bar{N}=\frac{N}{M}$, deci $n_j=\bar{N}\cdot f_2=\bar{n}$, atunci au loc relațiile:

$$T=N\cdot\bar{y}$$

$$s_{\hat{t}}^2=N^2\left(\frac{1-f_1}{m}\cdot\frac{1}{m-1}\sum_{j=1}^{m}(\bar{y}_j-\bar{y})^2+\frac{1-f_2}{m\cdot\bar{n}}\cdot\frac{1}{m}\sum_{j=1}^{m}s_j'^2\right)$$

$$s_j'^2=\frac{1}{\bar{n}-1}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2$$

$$\bar{y}_j=\frac{1}{\bar{n}}\sum_{i=1}^{n_j}y_{ij}$$

$$\bar{y}=\frac{1}{m}\sum_{j=1}^{m}\bar{y}_j$$

$$s_{\hat{\mu}}^2 = \frac{1-f_1}{m} \cdot \frac{1}{m-1} \sum_{j=1}^{m} (\bar{y}_j - \bar{y})^2 + f_1 \frac{1-f_2}{m \cdot \bar{n}} \cdot \frac{1}{m} \sum_{j=1}^{m} s_j'^2$$

2. Varianța estimatorului obținut prin metoda sondajului în două trepte este mai mare decât cea obținută cu ajutorul sondajului aleator simplu. Raportul dintre cele două variante se numește *efect de sondaj* (*design effect-deff*):

$$deff = \frac{V(\hat{\mu})}{V_{sas}(\hat{\mu})}$$

unde $V_{sas}(\hat{\mu})$ este varianța estimatorului mediei pentru cazul sondajului aleator simplu (SAS).

Efectul de sondaj se poate măsura cu ajutorul unui coeficient ce poartă numele de coeficient de corelație *intra-UP* notat cu ρ. Dacă UP au același volum $\bar{N}$, atunci $\sigma_b'^2 = \sigma_b^2$ și se poate arăta că:

$$\rho = \frac{\sigma_b^2}{\sigma^2} - \frac{1}{\bar{N}}$$

În aceste condiții, efectul de sondaj se obține astfel:

$$deff = 1 + \rho(\bar{n} - 1)$$

Ca o concluzie, pentru a diminua eroarea de estimare și a obține un rezultat apropiat de cel al sondajului aleator simplu, trebuie diminuată varianța *inter-UP* sau să crească componenta *intra-UP*. Cu alte cuvinte, gradul de omogenitate al UP nu ar trebui să fie semnificativ diferit de cel al populației totale.

3. Dacă se dispune de o valoare a lui ρ, atunci se poate calcula nivelul de precizie a unui sondaj în două trepte cu ajutorul relației:

$$V(\hat{t}) = N^2 \frac{\sigma^2}{m \cdot \bar{n}} (1 + \rho(\bar{n} - 1))$$

$$V(\hat{\mu}) = \frac{\sigma^2}{m \cdot \bar{n}} (1 + \rho(\bar{n} - 1))$$

În relația de mai sus, $m \cdot \bar{n}$ reprezintă volumul eșantionului extras cu ajutorul metodei sondajului în două trepte.

4. În condițiile în care se cunosc ρ și elementele de cost ale sondajului, adică costul total (C) și costurile unitare la nivel de UP (c_1) și de US (c_2), se pot estima efectivele m și $\bar{n}$, pentru o varianță $V(\hat{\mu})$ minimă, potrivit relațiilor de mai jos (Ardilly, 1994, p. 123):

$$\bar{n} = \sqrt{\frac{c_1}{c_2} \cdot \frac{1-\rho}{\rho}}$$

$$m = \frac{C}{c_1 + \sqrt{c_1 c_2} \cdot \frac{1-\rho}{\rho}}$$

$$C = c_1 m + c_2 m \cdot \bar{n}$$

Dacă se cunoaște numărul de unități ce trebuie extras dintr-o UP, eroarea maxim admisibilă cu care se estimează parametrul medie și anumite limite pentru varianțele *inter-UP*, atunci numărul de UP se poate obține astfel:

$$m = \frac{z^2}{\Delta_\mu^2} \frac{\sigma_b^2}{\bar{n}} = \frac{1}{SE^2} \frac{\sigma_b^2}{\bar{n}}$$

Exemplu

Considerăm populația fermelor producătoare de fructe dintr-o țară structurate în 40 de grupe ($M = 40$), după regiunea administrativă, și se cunoaște că fiecare regiune conține 100 de ferme ($N = 4000$). Pentru a studia nivelul producției de fructe (tone/ha), se realizează un sondaj în două trepte și se extrag aleator 5 regiuni ($m = 5$), iar apoi din fiecare regiune se extrag aleator simplu câte 20 de ferme ($n = 100$). Rezultatele obținute la nivel de eșantion sunt prezentate mai jos:

	1	**2**	**3**	**4**	**5**
n_j	20	20	20	20	20
N_j	100	100	100	100	100
$\bar{y}_j$	5	4,2	5,1	4,4	3,8
$s_j'^2$	1,69	1	1,44	1,21	1

Se cere să se estimeze punctual și prin interval de încredere nivelul mediu al producției agricole pe o firmă la nivelul întregii populații, cu o probabilitate de 0,95.

Se știe că:

$$N_j = \bar{N} = \frac{N}{M} = \frac{4000}{40} = 100$$

$$n_j = \bar{n} = \frac{n}{m} = \frac{100}{5} = 20$$

$$f_1 = \frac{m}{M} = \frac{5}{40} = 0{,}125$$

$$f_2 = \frac{n_j}{N_j} = \frac{20}{100} = 0{,}2$$

Nivelul mediu al producției de fructe pe o fermă la nivel de eșantion este:

$$\bar{y} = \frac{1}{m}\sum_{j=1}^{m} \bar{y}_j = 4{,}5 \text{ tone la ha}$$

Această valoare reprezintă estimația nedeplasată a mediei la nivelul populației totale.

Pentru o probabilitate de 0,95, intervalul de încredere este de forma:

$$\left(\bar{y} \pm 1{,}96 \cdot s_{\hat{\bar{\mu}}}\right)$$

unde

$$s^2_{\hat{\bar{\mu}}} = \frac{1-f_1}{m} \cdot \frac{1}{m-1}\sum_{j=1}^{m}\left(\bar{y}_j - \bar{y}\right)^2 + f_1 \frac{1-f_2}{m \cdot \bar{n}} \cdot \frac{1}{m}\sum_{j=1}^{m} s_j'^2$$

$$s^2_{\hat{\bar{\mu}}} = \frac{1-0{,}125}{5} \cdot \frac{1}{5-1} \cdot 1{,}2 + 0{,}125 \cdot \frac{1-0{,}2}{5 \cdot 20} \cdot \frac{1}{5} \cdot 6{,}34 = 0{,}053$$

$$s_{\hat{\bar{\mu}}} = 0{,}23 \text{ tone la ha}$$

Intervalul de încredere estimat este de forma:

$$\left(\bar{y} \pm 1{,}96 \cdot s_{\hat{\bar{\mu}}}\right) \Rightarrow (4{,}5 \pm 1{,}96 \cdot 0{,}23) \Rightarrow (4{,}5 \pm 0{,}45) \text{ tone la ha}$$

b. Metode de ameliorare a preciziei sondajului în două sau mai multe trepte

Există câteva posibilități de a îmbunătăți calitatea sondajului în două sau mai multe trepte.

Stratificarea unităților primare

Stratificarea introduce, de obicei, un câștig de precizie. Această procedură se poate aplica și în cazul sondajului în două trepte. Unitățile primare pot fi împărțite pe straturi după un anumit criteriu determinant pentru tema anchetei. De exemplu, dacă se consideră un sondaj despre nivelul investițiilor în România, unitățile primare sunt reprezentate de județele României, care sunt stratificate pe regiuni de dezvoltare economică. În interiorul fiecărei unități primare se găsesc baze de sondaj cu firmele din respectivul județ (existente la Camera de Comerț și Industrie sau la Direcția Finanțelor Publice), care reprezintă unitățile secundare. O unitate primară de dimensiuni mari, cum este municipiul București, trebuie considerată separat un strat. Recomandabil este ca straturile să fie de dimensiuni apropiate, atât ca număr de unități primare, cât și ca număr de unități secundare.

Alocarea unităţilor primare în eşantion, precum şi extragerea unităţilor secundare prezintă probleme practice complexe. Pentru simplificare, se poate aplica un sondaj stratificat cu alocare proporţională. Dacă informaţiile disponibile permit, se poate realiza şi o alocare optimală, în funcţie de cost şi de varianţa de la nivelul fiecărui strat. De exemplu, dacă varianţa şi costul unitar variază de la un strat la altul, este recomandabil ca fracţia de sondaj să fie proporţională cu varianţa din acel strat şi invers proporţională cu costul unitar din stratul respectiv.

Extragerea sistematică a unităţilor primare

Printr-o extragere sistematică se poate obţine un plus de precizie datorită modului de organizare a unităţilor care vor intra în eşantion.

Stratificarea unităţilor secundare în cadrul fiecărei unităţi primare

Precizia unui sondaj în două trepte poate fi ameliorată şi prin realizarea unui sondaj stratificat în treapta a doua a sondajului. După ce au fost extrase unităţile primare, extragerea unităţilor secundare se realizează printr-un procedeu specific sondajului stratificat. Baza de sondaj a fiecărei unităţi primare este structurată după o variabilă cunoscută. Extragerile din fiecare unitate primară sunt independente şi respectă condiţiile sondajului stratificat.

4.4. Sondajul pe serii (cluster)

În practică, există situaţii în care unităţile populaţiei sunt organizate în mod natural în grupuri sau în unităţi primare (UP) compacte numite serii sau clustere. De exemplu, vorbim despre persoanele dintr-o gospodărie sau dintr-un bloc de locuinţe; gospodăriile dintr-o zonă a unui oraş sau de pe o stradă; studenţii dintr-o grupă etc. O serie (cluster) reprezintă un număr de unităţi individuale din populaţie cu proprietatea că fiecare unitate aparţine unei singure serii.

4.4.1. Justificarea metodei

În principiu, acest tip de sondaj presupune ca unitatea de eşantionare să fie una complexă, adică o unitate formată din mai multe unităţi individuale ale populaţiei. Ca urmare, sondajul pe serii poate fi tratat ca un caz particular al sondajului în două trepte. Seriile identificate la nivelul populaţiei joacă rolul de UP şi reprezintă unităţile de eşantionare. Această metodă presupune ca în a doua treaptă a sondajului să nu se mai realizeze o extragere aleatoare de unităţi secundare (US), ci să se efectueze o cercetare exhaustivă a unităţilor individuale din fiecare UP extrasă în prima treaptă (vezi tabelul 4.3).

De exemplu, UP sau seriile pot fi gospodării de persoane, clădiri cu mai multe gospodării sau străzi cu mai multe clădiri. După ce au fost extrase în prima treaptă un număr de gospodării sau de clădiri, eşantionul supus observării este reprezentat de toate persoanele care compun aceste UP.

Avantajele metodei sunt în primul rând de ordin economic şi practic pentru că se reduc costurile de observare şi se simplifică procedura de constituire a bazei de sondaj. În plus, se reduc non-răspunsurile şi erorile de acoperire cauzate de incompletitudinea bazei de sondaj.

Inconvenientul metodei este dat de pierderea de precizie, efectul redundanţei informaţiei la nivel de serii putând fi unul ridicat. Pentru a reduce acest efect, seriile trebuie să îndeplinească următoarele caracteristici (Ardilly, 1994, p. 116):

- grupele să fie cât mai asemănătoare ca mărime şi structură;
- grupele să fie eterogene în raport cu variabila de interes;
- populaţia să fie structurată într-un număr suficient de mare de serii de unităţi individuale, ceea ce presupune ca volumul unei grupe să nu fie mare;
- în cadrul eşantionului să se extragă un număr mare de serii.

Tabelul 4.3. *Principiul selecţiei pe serii*

UP	**Etragere *UP* şi a eşantionului**
○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
○ ○ ○ ○ ○	● ● ● ● ●
○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
○ ○ ○ ○ ○ ○ ○ ○ ○ ○	● ● ● ● ● ● ● ● ● ●
...	...
○ ○ ○ ○ ○	● ● ● ● ●

4.4.2. Calitatea estimatorilor şi a rezultatelor

Considerăm mai întâi cazul unui sondaj în care seriile au acelaşi volum. Adică la nivelul populaţiei există M serii sau UP cu acelaşi volum:

$$N_j = \overline{N} = \frac{N}{M}$$

De asemenea, se consideră că se extrag din populaţie m serii de volum egal.

În comparaţie cu sondajul în două trepte, sondajul pe serii poate fi uşor comparat cu sondajul aleator simplu, doar că vorbim despre extragerea de UP şi nu de unităţi individuale.

În acest caz, avem relațiile:

– media și totalul la nivel de cluster:

$$\mu_j = \frac{1}{\bar{N}} \sum_{i=1}^{\bar{N}} y_{ij}$$

$$\tau_j = \sum_{i=1}^{\bar{N}} y_{ij}$$

– media la nivelul populației:

$$\mu = \frac{1}{N} \sum_{i=1}^{\bar{N}} \sum_{j=1}^{M} y_{ij}$$

unde

$$N = \bar{N} \cdot M$$

– estimatorul mediei:

$$\hat{\mu} = \frac{1}{m \cdot \bar{N}} \sum_{i=1}^{\bar{N}} \sum_{j=1}^{m} Y_{ij}$$

unde

$$n = m \cdot \bar{N}$$

și estimatorul este identic cu cel din cazul sondajului aleator simplu.
La nivel de eșantion, estimația mediei este:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{\bar{N}} \sum_{j=1}^{m} y_{ij}$$

– varianța estimatorului mediei:

$$V(\hat{\mu}) = \left(1 - \frac{m}{M}\right) \cdot \frac{1}{m} \cdot \frac{1}{(M-1)} \sum_{j=1}^{M} (\tau_j - \bar{\tau})^2$$

unde

$$\bar{\tau} = \frac{1}{M} \sum_{j=1}^{M} \tau_j$$

– estimația varianței estimatorului mediei:

$$s_{\hat{\mu}}^2 = \left(1 - \frac{m}{M}\right) \cdot \frac{1}{m} \cdot \frac{1}{(m-1)} \sum_{j=1}^{m} (T_j - \bar{T})^2$$

unde

$$\bar{T} = \frac{T}{m}$$

iar T este estimația punctuală a lui τ, respectiv T_j estimația lui τ_j.

– estimația varianței estimatorului totalului:

$$s_{\hat{\tau}}^2 = M^2 \cdot s_{\hat{\mu}}^2$$

Observații

– în comparație cu sondajul în două trepte, sondajul pe serii reține în relația varianței estimatorului mediei doar componenta varianței între UP, ceea ce înseamnă că este mai precis;

– efectul de sondaj $deff = \frac{V(\hat{\mu})}{V_{sas}(\hat{\mu})}$ ne arată că sondajul pe serii este mai precis decât cel aleator simplu dacă varianța dintre grupe nu este mare, adică dacă seriile sunt eterogene și varianța din interiorul lor preia o pondere mare din varianța totală.

– dacă seriile sau UP nu au același volum, formulele de mai sus se modifică cu ponderile corespunzătoare, după formulele de la sondajul în două trepte.

Exemplu

Pentru a estima nivelul mediu al producției de legume al unei gospodării (kg), se realizează un sondaj la nivelul celor 2000 de gospodării ale unei comune ce sunt organizate pe 50 de zone agricole a câte 40 de gospodării fiecare. În acest scop, se realizează un sondaj pe serii. Se extrag aleator 8 regiuni și se culeg date de la toate gospodăriile care formează astfel eșantionul.

Rezultatele obținute la nivelul eșantionului, privind producția totală de legume pentru gospodăriile din cele 8 regiuni extrase, sunt prezentate în tabelul de mai jos:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
N_j	40	40	40	40	40	40	40	40
T_j	12000	12300	13000	12500	12200	12100	12800	11500

Se cere să se estimeze punctual și prin interval de încredere nivelul mediu al producției de legume pe o gospodărie la nivelul întregii populații, cu o probabilitate de 0,95.

Se știe că:

$$N_j = \bar{N} = \frac{N}{M} = \frac{2000}{50} = 40$$

$$m = 8$$

$$f_1 = \frac{m}{M} = \frac{8}{50} = 0{,}16$$

$$n = m \cdot \bar{N} = 8 \cdot 40 = 320$$

Nivelul mediu al producției de fructe pe o fermă la nivel de eșantion este:

$$\bar{y} = \frac{T}{n} = \frac{1}{n}\sum_{j=1}^{m} T_j = \frac{1}{320} \cdot 98400 = 307{,}5 \text{ kg}$$

Această valoare reprezintă estimația nedeplasată a mediei la nivelul populației totale.

Pentru o probabilitate de 0,95, intervalul de încredere este de forma:

$$\left(\bar{y} \pm 1{,}96 \cdot s_{\hat{\bar{y}}}\right)$$

unde

$$s^2_{\hat{\bar{y}}} = \left(1 - \frac{m}{M}\right) \cdot \frac{1}{m(m-1)} \sum_{j=1}^{m} \left(T_j - \bar{T}\right)^2$$

iar

$$\bar{T} = \frac{T}{m}$$

$$T = \sum_{j=1}^{m} T_j = 98400$$

deci

$$\bar{T} = \frac{98400}{8} = 12300 \text{ kg}$$

$$s^2_{\hat{\bar{y}}} = \left(1 - \frac{8}{50}\right) \cdot \frac{1}{8 \cdot 7} \cdot 1560000 = 23400$$

$$s_{\hat{\bar{y}}} = 152{,}97 \text{ kg}$$

Intervalul de încredere estimat este de forma:

$$\left(\bar{y} \pm 1{,}96 \cdot s_{\hat{\bar{y}}}\right) \Rightarrow (307{,}5 \pm 1{,}96 \cdot 152{,}97) \Rightarrow (307{,}5 \pm 299{,}82) \text{ kg}$$

4.5. Exerciții

1. Se consideră o populație de volum $N = 12$, iar unitățile acesteia sunt organizate pe 4 straturi. Pentru o variabilă de interes, valorile pe cele 4 straturi se repartizează după cum urmează:

{2,4};
{1,3,5};
{1,2,4,4};
{3,4,5}.

Se cere:

- să se calculeze media și varianța variabilei de interes pe fiecare strat;
- să se verifice proprietățile mediei și ale varianței: media mediilor de strat este media populației totale; varianța totală se descompune în varianța între straturi și intra-straturi.

2. Pentru datele de la problema 1 se consideră extragerea fără revenire a unui eșantion de 4 unități, câte una din fiecare strat. Valorile la nivelul eșantionului sunt: {2,3,4,5}.

Se cere:

- să se estimeze punctual și prin interval de încredere media populației cu ajutorul eșantionului extras;
- să se calculeze eroarea efectivă de estimare și să se compare cu eroarea standard estimată.

3. La nivelul unui oraș ce conține 10000 de gospodării se organizează un sondaj stratificat pentru a estima proporția celor care dețin o locuință în proprietate privată. Se consideră că gospodăriile sunt structurate pe 4 straturi, care reprezintă cartierele orașului.

În urma eșantionării a 500 de gospodării, rezultă următoarele informații privind structura populației totale, structura eșantionului și ponderea celor care dețin locuința în proprietate:

	Cartier 1	**Cartier 2**	**Cartier 3**	**Cartier 4**
N_h	20%	25%	30%	25%
n_h	125	100	150	125
p_h	0,15	0,25	0,35	0,20

Se cere:

- să se estimeze punctual ponderea gospodăriilor care dețin în proprietate locuința. Ce proprietate are această estimație?
- să se estimeze prin interval de încredere, cu o probabilitate de 0,95, ponderea gospodăriilor care dețin în proprietate locuința.

4. La nivelul unui oraș cu 10000 de gospodării structurate în 4 cartiere se realizează un sondaj stratificat proporțional, de volum $n = 400$, pentru a estima nivelul mediu al venitului lunar pe gospodărie (euro). Rezultatele anchetei la nivel de eșantion, împreună cu structura populației pe cartiere sunt prezentate mai jos:

	Cartier 1	**Cartier 2**	**Cartier 3**	**Cartier 4**
N_h	2000	3000	2500	2500
$\bar{y}_h$	800	700	650	750
s'_h	100	75	60	80

Se cere:

- să se estimeze punctual nivelul mediu al venitului unei gospodării. Care este relația între această estimație și venitul mediu la nivel de eșantion?
- să se estimeze prin interval de încredere nivelul mediu al venitului unei gospodării, cu o probabilitate de 0,95;
- să se testeze dacă venitul mediu al unei gospodării nu diferă semnificativ de 700 de euro, admițând un risc de 5%.

5. Se cunoaște structura studenților unei universități pe 3 categorii de facultăți (socio-umane, științe, vocaționale) și se cunoaște, din studii anterioare, gradul de omogenitate al studenților de la fiecare facultate după gradul de participare la cursuri (numărul de prezențe la curs pe semestru pentru un student, cu valori între 0 și 14). Informațiile privind populația sunt prezentate în tabelul de mai jos:

	Categ. 1	**Categ. 2**	**Categ. 3**
N_h	1000	2500	1500
σ_h	2	3	2,5

Se dorește să se realizeze un sondaj stratificat de volum $n = 400$, pentru a studia gradul de participare a studenților la cursuri în anul curent. Se cere să se construiască structura eșantionului pentru o alocare proporțională și una optimală. Să se compare rezultatele.

6. Într-un sondaj statistic se dorește să se estimeze consumul mediu de gaze naturale pe lună al unei gospodării dintr-un cartier ($N = 5000$). Se consideră un plan de sondaj care ia în calcul organizarea gospodăriilor pe blocuri de apartamente. Într-o primă treaptă se extrag aleator 7 blocuri, iar în treapta a doua se includ în eșantion toate gospodăriile din aceste blocuri. La nivelul eșantionului s-au obținut rezultatele:

	1	**2**	**3**	**4**	**5**	**6**	**7**
Nr. gospod.	16	25	15	30	20	24	20
Consum total	200	320	190	400	250	300	240

Se cere:

- să se estimeze punctual valoarea totală a consumului și consumul mediu pe gospodărie la nivelul întregii populații;
- să se estimeze prin interval de încredere consumul total și nivelul mediu al consumului.

7. Pentru a estima nivelul mediu al venitului unei gospodării (euro) la nivelul unui oraș, se realizează un sondaj în două trepte. În prima treaptă se identifică un număr de 125 de străzi, cu un număr aproximativ egal de gospodării (100) din care se extrag aleator 10. În treapta a doua se extrag aleator 80 de gospodării de pe fiecare stradă.

Rezultatele la nivel de eșantion sunt prezentate mai jos:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
$\bar{y}_j$	1000	1250	1500	900	1400	950	1250	1100	1200	900
s'_j	100	110	160	100	120	80	95	110	100	70

Se cere să se estimeze punctual și prin interval de încredere nivelul mediu al venitului unei gospodării, cu o încredere de 0,95.

Capitolul 5

UTILIZAREA UNEI INFORMAȚII SUPLIMENTARE. AMELIORAREA CALITĂȚII ESTIMATORILOR

În realizarea sondajelor statistice, un rol important îl deține informația suplimentară cu privire la populația analizată. Dacă se dispune de o asemenea informație sub forma unei variabile puternic corelate cu variabila de interes, se pot construi planuri de sondaj mixte care pot îmbunătăți eficiența sondajului fie prin reducerea costurilor, fie prin ridicarea nivelului de precizie. De asemenea, informația suplimentară poate fi utilizată în etapa de estimare pentru a îmbunătăți eficiența estimatorilor specifici metodei de sondaj utilizate.

Sondajele aleatoare și mixte permit ca, în faza finală, după ce au fost culese datele statistice, să se intervină asupra acestora în sensul îmbunătățirii calității estimațiilor obținute. O asemenea intervenție presupune, de regulă, o ajustare cu ajutorul unei informații suplimentare care se utilizează a posteriori și care conduce la ameliorarea calității estimatorilor utilizați în inferența statistică. Calitatea unui estimator se poate îmbunătăți fie prin reducerea variației acestuia, fie a deplasării, dacă este cazul.

În acest capitol prezentăm trei posibilități de ameliorare a calității estimatorilor prin reducerea variației. În prima parte prezentăm principiul fundamental care stă la baza acestei proceduri. În următoarele trei secțiuni sunt prezentați estimatorii obținuți prin post-stratificare, raport și regresie.

5.1. Principiul ameliorării estimatorilor prin reducerea variației

Estimarea parametrilor unei populații la nivelul unui sondaj presupune utilizarea setului de date obținut cu ajutorul eșantionului extras. Pentru o variabilă de interes Y, acest eșantion este un set de date de forma:

$$(y_1, y_2, \dots, y_n)$$

Dacă se dispune de o informație suplimentară printr-o variabilă X strâns corelată cu variabila Y, atunci pentru fiecare unitate din eșantion se pot utiliza diverse măsuri obținute cu ajutorul unor instrumente statistice ce au la bază legătura dintre cele două variabile. Obiectivul final al acestui procedeu este acela de a obține un estimator pentru un parametru al variabilei de interes Y care să aibă o varianță mai redusă decât a estimatorului convențional (care nu ia în considerare variabila suplimentară X).

În general, informația suplimentară este disponibilă pentru toți indivizii populației, condiție esențială mai ales când este vorba despre construirea unui sondaj mixt. Practica arată însă că această informație este frecvent disponibilă sub forma unor indicatori (medie, varianță, proporție) și deci nu poate fi utilizată pentru a construi un plan de sondaj. În asemenea situații, informația poate fi utilă pentru a construi un alt estimator, după ce s-a realizat un sondaj aleator simplu.

Schematic, prezentăm procedeul de bază al ameliorării calității estimatorilor prin reducerea variației cu ajutorul etapelor de mai jos (Ardilly, 1994):

- se consideră o variabilă de interes Y la nivelul unei populații de referință P, care este studiată la nivelul unui sondaj s;
- după efectuarea sondajului, se determină un estimator pentru un parametru anume al distribuției populației după variabila Y, folosind o metodă clasică cunoscută (de exemplu, pentru parametrul medie se folosește estimatorul media de selecție);
- se stabilește o variabilă suplimentară X, puternic corelată cu variabila de interes Y; variabila X este cunoscută la nivelul întregii populații, adică este cunoscută distribuția acesteia (sunt cunoscuți parametrii distribuției); deținerea unei informații agregate nu presupune și cunoașterea valorilor variabilei la nivelul tuturor unităților individuale;
- la nivelul eșantionului, se determină un estimator cu varianță nulă al parametrului cunoscut al variabilei suplimentare X (de exemplu, dacă parametrul de estimat pentru variabila Y este media, atunci pentru variabila X se construiește un estimator tot pentru parametrul medie, numai că în acest caz parametrul este cunoscut, iar varianța estimatorului va fi nulă);
- se acționează asupra estimatorului inițial obținut pentru variabila Y în sensul modificării ponderilor care apar în expresia acestuia (pe baza ponderilor care apar în expresia estimatorului obținut pentru variabila X);
- se obține un nou estimator care pentru variabila X îndeplinește proprietatea cerută (varianță nulă), iar pentru variabila Y este un estimator mai bun, adică are o varianță mai mică.

Observație

Ameliorarea estimatorilor prin această schemă determină reducerea varianței, iar instrumentul de bază îl reprezintă modificarea ponderilor care apar în expresia estimatorului variabilei de interes la nivelul unui eșantion deja extras. Modificările care sunt realizate prin această metodă nu se realizează asupra eșantionului, adică asupra datelor deja culese.

5.2. Post-stratificarea

Prin post-stratificare se urmărește să se folosească proprietățile și precizia unui sondaj stratificat care să fie aplicate unui sondaj aleator simplu. În acest sens, se utilizează o variabilă suplimentară puternic corelată cu variabila de interes studiată. Conform principiului eficienței statistice, rezultatele acestei metode trebuie analizate din mai multe puncte de vedere: precizie, cost, simplitate, timp de realizare a procedeului.

Să presupunem că am extras un eșantion de volum n prin efectuarea unui sondaj aleator simplu. Având la dispoziție o variabilă suplimentară X, cunoscută la nivelul populației de referință, se face o stratificare a eșantionului extras în H straturi de volum n_h. Cele H straturi poartă numele de post-straturi.

De asemenea, presupunem că sunt cunoscute ponderile $\frac{N_h}{N}$ la nivelul colectivității totale, ponderi care dau stratificarea populației de referință după variabila X. Aceste ponderi îndeplinesc proprietatea:

$$N = \sum_{h=1}^{H} N_h$$

Estimatorul mediei, obținut pe baza informației suplimentare X, poartă numele de estimator prin post-stratificare și are relația:

$$\hat{\bar{Y}}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{\bar{y}}_h$$

unde $\hat{\bar{y}}_h$ reprezintă mediile de selecție de la nivelul fiecărui strat.

Aplicat variabilei X, acest estimator devine:

$$\hat{\bar{X}}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{\bar{x}}_h = \sum_{h=1}^{H} \frac{N_h}{N} \bar{X}_h = \bar{X}$$

adică îndeplinește proprietatea cerută de principiul menționat în capitolul introductiv.

Estimatorul obținut prin post-stratificare este, prin relație, identic cu cel obținut într-un sondaj stratificat, însă diferența majoră este următoarea:

- în cazul sondajului stratificat, ponderile $\frac{n_h}{n}$ de la nivelul eşantionului extras sunt stabilite prin metoda de eşantionare, proporţional sau nu;
- în cazul post-stratificării, eşantionul este obţinut aleator (după o metodă oarecare), iar volumul n_h este, de asemenea, aleator, el nu este stabilit în prealabil, ci doar are proprietatea:

$$\sum_{h=1}^{H} n_h = n$$

Se observă că prin post-stratificare am modificat ponderile estimatorului din sondajul iniţial. Înainte de a realiza post-stratificarea, estimatorul parametrului *medie* este media de selecţie:

$$\hat{\bar{y}} = \frac{1}{n}\sum_{i \in s} Y_i$$

iar ponderea sa este $1/n$.

După aplicarea metodei de ameliorare, expresia estimatorului este:

$$\hat{\bar{Y}}_{post} = \sum_{h=1}^{H}\sum_{i \in s_h} \frac{N_h}{N}\frac{1}{n_h} Y_i$$

iar ponderea sa are expresia

$$\frac{N_h}{N}\frac{1}{n_h}$$

Notând prin w_i ponderea care apare în expresia unui estimator, putem scrie:

$$w_i = \frac{1}{n} = constant$$

pentru estimatorul obţinut în mod clasic;

$$w_i(s) = \frac{N_h}{N}\frac{1}{n_h}$$

pentru estimatorul obţinut prin post-stratificare care depinde de n_h.

5.2.1. Proprietăţile estimatorului

Relaţia estimatorului obţinut prin post-stratificare este următoarea:

$$\hat{\bar{Y}}_{post} = \sum_{h=1}^{H} \frac{N_h}{N}\hat{\bar{y}}_h = \sum_{h=1}^{H}\sum_{i \in s_h} \frac{N_h}{N}\frac{1}{n_h} y_i$$

Verificăm proprietățile pentru acest estimator.

a. Nedeplasarea estimatorului

Proprietatea de nedeplasare este verificată prin relația:

$$E\left(\hat{\bar{Y}}_{post}\right) = E\left[\sum_{h=1}^{H} \frac{N_h}{N} \hat{\bar{y}}_h\right] = \sum_{h=1}^{H} \frac{N_h}{N} E(\hat{\bar{y}}_h) = \sum_{h=1}^{H} \frac{N_h}{N} \bar{Y} = \bar{Y}$$

Din rezultatul de mai sus rezultă că estimatorul obținut prin post-stratificare este nedeplasat, adică aplicarea acestei metode de redresare nu afectează proprietatea de nedeplasare a estimatorului inițial.

b. Varianța estimatorului

Proprietatea fundamentală a redresării este obținerea unui estimator care are o varianță mai mică decât a estimatorului determinat prin modelul inițial.

Varianța estimatorului obținut prin post-stratificare se determină doar pe baza varianței de la nivelul fiecărui substrat al eșantionului, adică componenta intra-strat. Deoarece estimatorul depinde de n_h, care este o variabilă aleatoare și care apare la numitor, determinarea varianței acestuia este mai complexă.

În calcule se folosește o dezvoltare în serie aproximată până la gradul doi a raportului $\frac{1}{n_h}$, pentru care este necesar să se aplice media. Varianța estimatorului are la bază relația:

$$V\left(\hat{\bar{Y}}_{post}\right) = V\left[\sum_{h=1}^{H} \frac{N_h}{N} \hat{\bar{y}}_h\right] = \sum_{h=1}^{H} \frac{N_h^2}{N^2} V(\hat{\bar{y}}_h)$$

Deoarece $\hat{\bar{y}}_h$ depinde de n_h, care este o variabilă aleatoare, atunci are loc relația:

$$V\left(\hat{\bar{Y}}_{post}\right) = \sum_{h=1}^{H} \frac{N_h^2}{N^2} V(\hat{\bar{y}}_h) = E\left[\sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h} \frac{1}{n_h} \sigma_h^2\right]$$

de unde rezultă

$$V\left(\hat{\bar{Y}}_{post}\right) = E\left[\sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{1}{n_h} \sigma_h^2 - \frac{1}{N} \sum_{h=1}^{H} \frac{N_h}{N} \sigma_h^2\right]$$

unde σ_h^2 reprezintă varianțele de la nivelul fiecărui strat.

În relația de mai sus ne interesează media raportului $\frac{1}{n_h}$. Pentru a calcula această medie vom folosi o dezvoltare în serie plecând de la relația:

$$n_h = \frac{n_h}{n} n = np_h = n(p_h - P_h + P_h) = nP_h\left(1 + \frac{p_h - P_h}{P_h}\right)$$

adică

$$\frac{1}{n_h} = \frac{1}{nP_h} \frac{1}{\left(1 + \frac{p_h - P_h}{P_h}\right)}$$

unde P_h este estimatorul lui p_h.

Dezvoltând în serie şi aproximând pentru primii doi termeni, rezultă:

$$\frac{1}{n_h} \cong \frac{1}{nP_h}\left(1 - \frac{p_h - P_h}{P_h} + \left(\frac{p_h - P_h}{P_h}\right)^2\right)$$

Aplicând media relaţiei de mai sus, obţinem:

$$E\left(\frac{1}{n_h}\right) \cong \frac{1}{nP_h}\left(1 + E\left(\frac{p_h - P_h}{P_h}\right)^2\right) = \frac{1}{nP_h}\left(1 + \frac{1}{P_h^2} V(p_h)\right)$$

$$E\left(\frac{1}{n_h}\right) = \frac{1}{nP_h}\left(1 + \frac{N-n}{n}\frac{1}{n}\frac{1}{P_h^2}\frac{N-N_h}{N}P_h\right) = \frac{1}{n}\frac{N}{N_h} + \frac{N-n}{n^2}\frac{N-N_h}{N_h}\frac{N}{N_h}$$

unde

$$P_h = \frac{N_h}{N}$$

Înlocuind în relaţia varianţei şi reducând termenii, obţinem expresia finală a varianţei estimatorului prin post-stratificare:

$$V\left(\hat{\bar{Y}}_{post}\right) = \frac{1-f}{n}\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2 + \frac{1-f}{n^2}\sum_{h=1}^{H}\frac{N-N_h}{N}\sigma_h^2$$

unde

$$\frac{1-f}{n}\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2$$

reprezintă varianţa estimatorului unei medii în sondajul stratificat proporţional, iar

$$\frac{1-f}{n^2}\sum_{h=1}^{H}\frac{N-N_h}{N}\sigma_h^2$$

este un factor care mărește varianța față de cazul stratificării proporționale. Se observă că acest termen este neglijabil pentru sondaje de volum mare.

Pentru calculul intervalelor de încredere, la nivelul unui eșantion, se utilizează următoarea estimație a variației estimatorului:

$$s^2_{\hat{\bar{Y}}_{post}} = \frac{1-f}{n}\sum_{h=1}^{H}\frac{N_h}{N}s_h'^2 + \frac{1-f}{n^2}\sum_{h=1}^{H}\frac{N-N_h}{N}s_h'^2$$

unde $s_h'^2$ sunt estimațiile variațelor straturilor, calculate la nivel de eșantion.

Observație

Deși precizia estimatorului obținut în urma redresării este cu puțin mai scăzută decât în cazul stratificării apriori, totuși se realizează un câștig de precizie față de sondajul aleator simplu. Metoda de redresare prin post-stratificare are avantajul că informația suplimentară nu trebuie cunoscută la nivelul fiecărei unități a colectivității, ci trebuie cunoscute doar valorile globale din fiecare strat N_h. Este evident că în această situație costul anchetei se reduce.

Din această observație rezultă imediat faptul că este preferabilă post-stratificarea în cazul în care dorim să diminuăm costul unei anchete, dacă ar fi să ne raportăm la un sondaj stratificat, iar dacă ar fi să ne raportăm la un sondaj aleator simplu, câștigul adus de post-stratificare este evident în orice caz.

5.2.2. Limitele și eficiența metodei

Metoda se poate aplica doar atunci când se cunosc, pe baza variabilei auxiliare X, ponderile $\frac{N_h}{N}$. Dacă aceste ponderi nu se cunosc, metoda nu se mai poate aplica.

Post-stratificarea devine ineficientă când post-straturile sunt prea numeroase. În practică, se recomandă ca $\frac{N_h}{N} \geq 10\%$.

Deoarece n_h este variabilă aleatoare, apare problema când n_h ia valoarea egală cu unu, caz în care nu se poate estima varianța σ_h^2. În cazul în care n_h ia valoarea unu, se poate face o corecție prin înlăturarea unității care apare în acel substrat.

În concluzie, metoda de redresare prin post-stratificare conduce la un estimator care, din punctul de vedere al eficienței statistice, este mai bun: pe de o parte, dacă ne raportăm la un sondaj aleator simplu, estimatorul prin post-stratificare are o varianță mai mică, adică este eficient, iar dacă ne raportăm la un sondaj stratificat, metoda ne oferă un estimator cu o varianță puțin mai mare, cantitate neglijabilă pentru eșantioane mari, însă implică costuri substanțial reduse.

Limitele de eficiență ale unui sondaj aleator pot fi îndepărtate doar parțial și în anumite condiții prin organizarea sondajului stratificat sau a sondajului în mai

multe trepte. Prin utilizarea unei informaţii suplimentare se poate îmbunătăţi nu numai precizia estimatorilor, ci şi alte elemente ale eficienţei unui sondaj, cum ar fi costul, timpul de realizare, dificultăţile practice, posibilitatea de a atinge obiectivele.

Prin post-stratificare se poate răspunde următoarelor limite ale unui sondaj organizat de manieră clasică: aleator, stratificat sau în mai multe trepte:

1. Riscul distorsiunilor de eşantionare (supra-reprezentarea sau sub-reprezentarea unor anumite categorii de unităţi din populaţie), care poate conduce la o precizie scăzută a rezultatelor. Această problemă este rezolvată de sondajul stratificat, însă în unele cazuri un astfel de sondaj nu se poate aplica.

2. Cunoaşterea variabilei suplimentare la nivelul întregii populaţii, pentru a realiza apartenenţa fiecărei unităţi individuale la un strat. Atât din punct de vedere practic (variabila X nu este cuprinsă în baza de sondaj sau variaţiile valorilor sale sunt prea mari), cât şi economic (este costisitor şi dificil de sortat baza de sondaj în funcţie de variabila suplimentară), această condiţie poate fi o restricţie deosebită.

Prin post-stratificare se pot corecta distorsiunile de eşantionare, iar precizia rezultatelor obţinute este cu puţin inferioară sondajului stratificat, însă costurile sunt mult mai reduse. Post-stratificarea nu impune decât cunoaşterea ponderilor de la nivelul fiecărui strat, condiţie mult mai facil de îndeplinit decât cunoaşterea variabilei la nivelul întregii populaţii.

3. Apariţia non-răspunsurilor, fie totale, fie parţiale. Aceste erori pot fi corectate prin supradimensionarea eşantionului, prin revizitarea unităţilor care lipsesc sau prin organizarea unei anchete de dimensiuni mai mici în cazul refuzurilor. Metoda post-stratificării, însă, oferă o soluţie eficientă în cazul corectării non-răspunsurilor. Utilizând variabile de control prin care se realizează stratificarea, non-răspunsurile pot fi corectate prin reponderare, la nivelul fiecărui strat. Metoda este suficient de precisă şi nu implică costuri ridicate.

Exemplu

Considerăm o anchetă asupra consumului de carne în oraşul Iaşi (date convenţionale), care îşi propune, printre altele, să estimeze consumul mediu lunar de carne pe un locuitor.

Rezultatele anchetei, ţinând cont de grupa de vârstă a populaţiei cuprinse în eşantion, în urma aplicării unui sondaj aleator simplu, sunt prezentate în tabelul 5.1. Pentru a îmbunătăţi calitatea estimatorului parametrului consum mediu de carne, ţinând cont de variabila suplimentară X - grupa de vârstă a populaţiei oraşului, se aplică procedeul de înlocuire a ponderilor obţinute la nivelul eşantionului cu cele oferite prin variabila suplimentară X.

Tabelul 5.1. *Rezultatele anchetei*

Grupa de vârstă X (ani)	**$\bar{y}_h$ consum mediu lunar de carne (kg)**	**$\frac{n_h}{n}$**
0 - 20	0,54	12%
21 - 35	0,75	27%
36 - 50	0,92	36%
50 și peste	1,12	25%
Total	-	**100%**

Conform tabelului 5.1, din eșantion au rezultat proporțiile $\frac{n_h}{n}$, care structurează eșantionul în raport cu variabila grupa de vârstă. La nivelul eșantionului, se obțin patru post-straturi.

Consumul mediu lunar de carne pe un locuitor este estimat punctual cu ajutorul mediei de la nivelul eșantionului. Ținând cont de cele patru post-straturi și conform rezultatelor din tabelul 5.1, se calculează consumul mediu la nivelul eșantionului prin relația:

$$\bar{y} = \sum_{h=1}^{4} \frac{n_h}{n} \bar{y}_h = 0{,}54 \cdot 0{,}12 + 0{,}75 \cdot 0{,}27 + 0{,}92 \cdot 0{,}36 + 1{,}12 \cdot 0{,}25$$

$$\bar{y} = 0{,}879 \text{ kg}$$

Așadar, consumul mediu lunar de carne pe un locuitor este estimat punctual, pe baza estimației obținute la nivelul eșantionului, la 0,879 kg carne.

La nivelul populației de referință (populația orașului Iași), sunt cunoscute proporțiile pe grupe de vârstă, conform tabelului 5.2.

Din tabelul 5.2 se observă că ponderile din eșantion diferă de cele de la nivelul populației (pentru persoanele de peste 50 de ani sunt mai mici și pentru persoanele de sub 20 de ani sunt mai mari).

Tabelul 5.2. *Structura populației orașului Iași pe grupe de vârstă*

Grupa de vârstă X (ani)	**$\frac{N_h}{N}$**
0 - 20	15%
21 - 35	22%
36 - 50	33%
50 și peste	30%
Total	**100%**

Estimatorul obținut prin post-stratificare se determină înlocuind ponderile de la nivelul eșantionului $\frac{n_h}{n}$ cu cele cunoscute pe baza variabilei X, respectiv ponderile $\frac{N_h}{N}$.

Pe baza estimatorului obţinut prin post-stratificare, se obţine o estimaţie mai bună:

$$\bar{y}_{post} = \sum_{h=1}^{4} \frac{N_h}{N} \bar{y}_h = 0{,}54 \cdot 0{,}15 + 0{,}75 \cdot 0{,}22 + 0{,}92 \cdot 0{,}33 + 1{,}12 \cdot 0{,}30$$

$$\bar{y}_{post} = 0{,}885 \text{ kg}$$

Prin post-stratificare se obţine o valoare a estimatorului mai mare şi mai reală. Atât $\bar{y}$ cât şi $\bar{y}_{post}$ sunt estimaţii obţinute cu ajutorul unor estimatori nedeplasaţi ai mediei, dar $\bar{y}_{post}$ este mai precis.

5.3. Estimatorul raport (*ratio*)

În practică, frecvent este foarte important să se cunoască raportul a doi parametri (medii sau totaluri). De exemplu, în anchetele economice este important să se estimeze venitul pe locuitor, rata şomajului etc. Pentru asemenea situaţii, se construieşte un estimator care are la bază un raport între doi parametri obţinuţi pentru două variabile. Ideea de bază a metodei este aceea de a dispune de informaţia suplimentară la nivelul întregii populaţii, adică de a cunoaşte unul dintre parametri. Aşadar, metoda se foloseşte în cazurile în care se cunoaşte o variabilă suplimentară (X) corelată cu variabila de interes (Y) şi pentru care există raţiuni teoretice şi practice pentru a estima un raport de forma:

$$r_i = \frac{y_i}{x_i}$$

5.3.1. Principiu

Înainte de a intra în detaliile metodei, este important de subliniat că acest principiu a fost utilizat în mod concret de Laplace, în contextul unei anchete ce îşi propune să estimeze populaţia Franţei în anul 1802. Rezultatele au fost publicate în lucrarea *Essai philosophique sur les probabilités*, în anul 1814. Laplace construieşte un eşantion de 30 de comune franceze şi obţine un total al populaţiei de 2037615 de persoane. Pentru aceste comune, autorul cunoaşte numărul de naşteri, care este egal cu 71866, şi poate estima raportul dintre numărul populaţiei şi numărul de naşteri:

$$r = \frac{y}{x} = 28{,}35$$

De asemenea, Laplace consideră că dacă are la dispoziţie informaţii cu privire la numărul total de naşteri în anul 1802, poate construi o estimaţie pentru populaţia Franţei folosind regula de trei simplă:

$$total\ locuitori = total\ nașteri \cdot \frac{total\ locuitori\ la\ nivel\ de\ eșantion}{total\ nașteri\ la\ nivel\ de\ eșantion}$$

În exemplul lui Laplace, informația suplimentară este reprezentată de numărul total de nașteri la nivelul populației. Această informație este puternic corelată cu numărul de locuitori. Cu alte cuvinte, în acest exemplu se lucrează cu parametrul total, iar ca informație suplimentară se dispune de acest parametru la nivelul populației. De asemenea, calculul realizat de Laplace presupune și utilizarea unei estimații a parametrului la nivel de eșantion. În final, autorul dispune de o informație care poate fi calculată și ca un raport între parametru și estimație. Pentru toate eșantioanele posibil de extras, aceasta este un raport între parametru și estimator:

$$\frac{\tau_X}{\hat{\tau}_X}$$

sau

$$\frac{\bar{X}}{\hat{\bar{x}}}$$

Pentru a finaliza exemplul, Laplace consideră cunoscut numărul de nașteri în Franța, în 1802, aproximându-l la un milion de persoane, și conchide că estimația cu privire la întreaga populație este obținută multiplicând această valoare cu raportul calculat mai sus (egal cu 28,35) și obține o populație totală de aproximativ 28,35 de milioane de persoane la nivelul Franței.

Se consideră că variabila X, variabila suplimentară folosită în procedeul de ameliorare, este legată de variabila de interes Y printr-o relație de forma:

$$y_i = R \cdot x_i + u_i$$

În relația de mai sus, R este un coeficient de proporționalitate, un raport, iar u_i sunt valori, numite reziduuri, care descriu abaterile de proporționalitate dintre cele două variabile X și Y. Metoda se folosește atunci când valorile u_i sunt mici.

Pentru reziduuri se cunoaște proprietatea

$$\sum_{i=1}^{N} u_i = 0$$

adică acestea se compensează la nivelul populației de referință, dacă și numai dacă are loc relația:

$$R = \frac{\bar{Y}}{\bar{X}}$$

În urma extragerii unui eşantion reprezentativ de volum n, într-un sondaj aleator simplu, de exemplu, se poate considera că şi la nivelul acestuia se compensează reciproc valorile reziduale, adică $\sum_{i=1}^{n} \hat{u}_i = 0$, ceea ce este echivalent cu a identifica o estimaţie a coeficientului de proporţionalitate cu ajutorul relaţiei de forma:

$$R' = \frac{\bar{y}}{\bar{x}}$$

Plecând de la acest principiu, se poate construi un estimator pentru media variabilei de interes prin relaţia:

$$\hat{\bar{Y}}_R = \frac{\hat{\bar{y}}}{\hat{\bar{x}}} \cdot \bar{X}$$

sau

$$\hat{\bar{Y}}_R = \hat{\bar{y}} \cdot \frac{\bar{X}}{\hat{\bar{x}}}$$

sau

$$\hat{\bar{Y}}_R = \hat{R} \cdot \bar{X}$$

unde

$$\hat{R} = \frac{\hat{\bar{y}}}{\hat{\bar{x}}}$$

Estimatorul $\hat{\bar{Y}}_R$ poartă numele de estimator raport.

În cazul variabilei X, avem:

$$\bar{X}_R = \hat{\bar{x}} \cdot \frac{\bar{X}}{\hat{\bar{x}}} = \bar{X}$$

adică estimatorul raport îndeplineşte condiţia fundamentală prezentată în capitolul introductiv cu privire la tehnica ameliorării (estimatorul estimează cu varianţă zero parametrul pentru variabila X).

Observaţia 1

În relaţia estimatorului raport apare, pe lângă estimatorul iniţial al mediei, un coeficient, un raport dintre media variabilei suplimentare şi estimatorul acesteia. La numitorul relaţiei este o variabilă aleatoare, estimatorul mediei variabilei X. Estimatorul raport este obţinut ca raport a două variabile aleatoare, a doi estimatori: estimatorul iniţial şi estimatorul determinat de variabila suplimentară X (media acesteia este cunoscută).

Observația 2

Estimatorul construit mai sus este valabil și pentru parametrul total, pentru că raportul dintre două medii este egal cu raportul dintre două totaluri:

$$\frac{\mu_Y}{\mu_X} = \frac{\tau_Y}{\tau_X}$$

deci

$$\hat{\tau}_Y = \frac{\bar{y}}{\bar{x}} \cdot \tau_X$$

5.3.2. Proprietățile estimatorului

Prezentăm proprietățile acestui estimator: nedeplasarea și eficiența.

a. Proprietatea de nedeplasare

Deoarece la numitorul estimatorului apare o variabilă aleatoare, pentru a putea aplica media se foloseşte o aproximare de ordinul unu în dezvoltarea în serie a estimatorului, în ipoteza că volumul eşantionului este mare şi că termenii în $1/n^2$ sunt neglijabili.

În aceste condiţii se poate scrie:

$$\hat{\bar{Y}}_R = \hat{\bar{y}} \cdot \frac{\bar{X}}{\hat{\bar{x}}} = \bar{X} \cdot \frac{\hat{\bar{y}} - \bar{Y} + \bar{Y}}{\hat{\bar{x}} - \bar{X} + \bar{X}} = \bar{Y} \cdot \frac{1 + \frac{\hat{\bar{y}} - \bar{Y}}{\bar{Y}}}{1 + \frac{\hat{\bar{x}} - \bar{X}}{\bar{X}}}$$

iar prin dezvoltarea în serie obţinem:

$$\hat{\bar{Y}}_R = \bar{Y}\left[\left(1 + \frac{\hat{\bar{y}} - \bar{Y}}{\bar{Y}}\right)\left(1 - \frac{\hat{\bar{x}} - \bar{X}}{\bar{X}} + \left(\frac{\hat{\bar{x}} - \bar{X}}{\bar{X}}\right)^2 - \cdots\right)\right]$$

Prin aproximare, pentru un ordin de mărime în $1/n$, se obţine:

$$\hat{\bar{Y}}_R \cong \bar{Y}\left(1 + \frac{\hat{\bar{y}} - \bar{Y}}{\bar{Y}} - \frac{\hat{\bar{x}} - \bar{X}}{\bar{X}} - \frac{(\hat{\bar{x}} - \bar{X})(\hat{\bar{y}} - \bar{Y})}{\bar{X} \cdot \bar{Y}} + \left(\frac{\hat{\bar{x}} - \bar{X}}{\bar{X}}\right)^2 + \left(\frac{\hat{\bar{y}} - \bar{Y}}{\bar{Y}}\right)\left(\frac{\hat{\bar{x}} - \bar{X}}{\bar{X}}\right)^2\right)$$

Aplicând media expresiei de mai sus, avem:

$$E\left(\hat{\bar{Y}}_R\right) = \bar{Y}\left[1 - \frac{cov(\hat{\bar{x}}, \hat{\bar{y}})}{\bar{X} \cdot \bar{Y}} + \frac{V(\hat{\bar{x}})}{\bar{X}^2}\right]$$

Din relaţia de mai sus se observă că estimatorul raport este uşor deplasat.

Dacă explicităm relaţia de mai sus folosind relaţiile:

$$V(\hat{\bar{x}}) = \frac{1}{n}V(X) = \frac{1}{n}\sigma_X^2$$

$$cov(\hat{\bar{x}}, \hat{\bar{y}}) = \frac{1}{n}\rho(X,Y) \cdot \sigma_X \cdot \sigma_Y$$

obţinem:

$$E\left(\hat{\bar{Y}}_R\right) = \bar{Y}\left[1 - \frac{1}{n}\frac{\rho(X,Y) \cdot \sigma_X \cdot \sigma_Y}{\bar{X} \cdot \bar{Y}} + \frac{1}{n}\frac{\sigma_X^2}{\bar{X}^2}\right]$$

Deplasarea estimatorului este:

$$B(n, \bar{Y}) = \frac{1}{n}\bar{Y}\left[\frac{\sigma_X^2}{\bar{X}^2} - \frac{\rho(X,Y) \cdot \sigma_X \cdot \sigma_Y}{\bar{X} \cdot \bar{Y}}\right]$$

care poate fi considerată nesemnificativă pentru un volum mare al eşantionului.

Observaţii

- dacă coeficientul de corelaţie dintre variabilele X şi Y este pozitiv, adică între cele două variabile există o legătură directă, atunci deplasarea este mică;
- deplasarea este nulă atunci când

$$\frac{\rho(X,Y) \cdot \sigma_Y}{\sigma_X} = \frac{\bar{Y}}{\bar{X}}$$

sau

$$\bar{Y} = \bar{X} \cdot \frac{\rho(X,Y) \cdot \sigma_Y}{\sigma_X}$$

adică atunci când variaţia lui Y este proporţională cu cea a lui X (dreapta de regresie trece prin origine).

b. Varianţa estimatorului

Reamintim că expresia estimatorului raport este:

$$\hat{\bar{Y}}_R = \hat{\bar{y}} \cdot \frac{\bar{X}}{\hat{\bar{x}}}$$

Pentru determinarea varianţei, se va realiza o dezvoltare a diferenţei:

$$\hat{\bar{Y}}_R - \bar{Y} = \hat{\bar{y}} \cdot \frac{\bar{X}}{\hat{\bar{x}}} - \bar{Y}$$

Rezultă:

$$\hat{\bar{Y}}_R - \bar{Y} \cong \bar{Y}\left(\frac{\hat{\bar{y}} - \bar{Y}}{\bar{Y}} - \frac{\hat{\bar{x}} - \bar{X}}{\bar{X}} - \frac{(\hat{\bar{x}} - \bar{X})(\hat{\bar{y}} - \bar{Y})}{\bar{X} \cdot \bar{Y}} + \left(\frac{\hat{\bar{x}} - \bar{X}}{\bar{X}}\right)^2 + \left(\frac{\hat{\bar{y}} - \bar{Y}}{\bar{Y}}\right)\left(\frac{\hat{\bar{x}} - \bar{X}}{\bar{X}}\right)^2\right)$$

iar

$$\left(\hat{\bar{Y}}_R - \bar{Y}\right)^2 \cong \bar{Y}^2\left(\left(\frac{\hat{\bar{y}} - \bar{Y}}{\bar{Y}}\right)^2 - 2\frac{(\hat{\bar{x}} - \bar{X})(\hat{\bar{y}} - \bar{Y})}{\bar{X} \cdot \bar{Y}} + \left(\frac{\hat{\bar{x}} - \bar{X}}{\bar{X}}\right)^2\right)$$

În aceste condiţii,

$$V\left(\hat{\bar{Y}}_R\right) = M\left(\hat{\bar{Y}}_R - \bar{Y}\right)^2 \cong \bar{Y}^2\left(\frac{1}{\bar{Y}^2}V(\hat{\bar{y}}) - 2\frac{cov(\hat{\bar{x}}, \hat{\bar{y}})}{\bar{X} \cdot \bar{Y}} + \frac{1}{\bar{X}^2}V(\hat{\bar{x}})\right)$$

Notând cu

$$R = \frac{\bar{Y}}{\bar{X}}$$

avem:

$$V\left(\hat{\bar{Y}}_R\right) = \frac{1}{n}(\sigma_Y^2 - 2R \cdot \rho(X,Y) \cdot \sigma_X\sigma_Y + R^2 \cdot \sigma_X^2)$$

Dacă ţinem cont de reziduu $u_i = y_i - R \cdot x_i$, atunci varianţa estimatorului raport se poate scrie:

$$V\left(\hat{\bar{Y}}_R\right) = \frac{1}{n}\sigma_u^2$$

Pentru calculul unui interval de încredere, se poate estima varianța estimatorului raport cu ajutorul relației:

$$s^2_{\hat{\bar{Y}}_R} = \frac{1}{n}\left(s_Y'^2 - 2R' \cdot r_{xy} \cdot s_X's_Y' + R'^2 \cdot s_X'^2\right)$$

sau

$$s^2_{\hat{\bar{Y}}_R} = \frac{1}{n-1}s_u'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - R' \cdot x_i)^2$$

unde

$$R' = \frac{\bar{y}}{\bar{x}}$$

Comparând varianţa estimatorului raport cu varianţa estimatorului iniţial, obţinut într-un sondaj aleator simplu, se observă că se obţine un câştig de precizie dacă:

$$\frac{\sigma_Y}{\sigma_X}\rho(X,Y) > \frac{R}{2}$$

În concluzie, estimatorul raport este mai bun decât cel clasic dacă reziduurile au valori mici (au o varianță mică) şi dacă între variabila de interes Y şi variabila suplimentară X există o legătură aproximativ liniară, sub forma unei drepte care trece prin origine.

Exemplu

Considerăm un sondaj care are drept obiectiv estimarea venitului mediu al unui salariat din sectorul comercial dintr-un oraş (date convenționale).

Pe baza unui sondaj aleator simplu, venitul mediu lunar al unui salariat din comerţ s-a estimat punctual la o valoare de 2000 de lei. Cunoscând că la nivelul eşantionului s-a înregistrat un timp mediu de lucru zilnic al unei persoane de 9,2 ore şi cunoscând dintr-un studiu anterior că timpul mediu zilnic de lucru al unei persoane din sectorul comercial este de 9,8 ore, se poate redresa calitatea rezultatului obţinut cu ajutorul estimatorului raport, luând în considerare timpul de lucru al unui salariat din acest sector.

Notăm cu Y variabila venit lunar pe un salariat din comerţul oraşului considerat, iar cu X variabila timp de lucru zilnic pe un salariat. Se cunosc următoarele date:

$\bar{y} = 2000$ de lei, $\bar{x} = 9{,}2$ h, $\bar{X} = 9{,}8$ h.

Estimatorul raport are următoarea relaţie:

$$\hat{\bar{Y}}_R = \hat{\bar{y}} \cdot \frac{\bar{X}}{\hat{\bar{x}}}$$

Pentru datele de la nivelul eşantionului se obţine estimaţia:

$$\bar{y}_R = \bar{y} \cdot \frac{\bar{X}}{\bar{x}} = 2000 \cdot \frac{9{,}8}{9{,}2} = 2130{,}43 \text{ de lei}$$

Pe baza estimatorului raport, s-a obţinut o valoare estimată a venitului mediu mai mare decât în cazul estimatorului clasic, folosind media de selecţie. Acest rezultat este afectat de variabila timp de lucru zilnic, variabilă puternic corelată cu variabila de interes şi cu care se află într-un raport de proporţionalitate.

Sub aspectul eficienţei, metoda de redresare cu ajutorul estimatorului raport aduce nu numai un câştig relativ de precizie. Metoda poate fi comparată cu

sondajul aleator cu probabilități inegale, care însă este mult mai dificil de aplicat și mai costisitor.

5.4. Estimatorul regresie

Ca și în cazul metodei de ameliorare prezentate în capitolul anterior, se pornește de la premisa existenței unei variabile suplimentare X, puternic corelată cu variabila de interes, și că între aceste variabile există o dependență liniară.

5.4.1. Principiu

Această metodă de ameliorare are la bază următoarea relație între variabila de interes și variabila suplimentară:

$$y_i = \alpha + \beta \cdot x_i + u_i$$

Relația de mai sus reprezintă ecuația unei drepte care este afectată de reziduurile u_i. Alegerea parametrilor α, β se face prin îndeplinirea condiției de compensare a reziduurilor la nivelul populației.

Conform metodei celor mai mici pătrate, se obțin următoarele valori pentru parametrii α, β, care îndeplinesc condiția cerută mai sus:

$$\beta = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

și

$$\alpha = \bar{Y} - \beta\bar{X}$$

În aceste condiții, are loc:

$$\bar{Y} = \alpha + \beta\bar{X}$$

Dacă dorim să estimăm parametrul medie în condițiile unui sondaj aleator simplu, atunci se poate considera că pentru un eșantion reprezentativ se poate scrie:

$$\bar{y} = \alpha + \beta \cdot \bar{x}$$

Din relațiile de mai sus, se obține:

$$\bar{Y} - \bar{y} = \beta(\bar{X} - \bar{x})$$

sau

$$\bar{Y} = \bar{y} + \beta(\bar{X} - \bar{x})$$

Acest principiu a condus la construirea unui estimator pentru parametrul medie a variabilei Y, ținând cont de trei estimatori: estimatorul inițial al mediei, estimatorul mediei variabilei X și estimatorul parametrului β.

Astfel, se defineşte estimatorul:

$$\hat{\bar{Y}}_{reg} = \hat{\bar{y}} + \hat{\beta}(\bar{X} - \hat{\bar{x}})$$

care se numeşte *estimator regresie*, în care

$$\hat{\beta} = \frac{\sum_{i \in s}(x_i - \hat{\bar{x}})(y_i - \hat{\bar{y}})}{\sum_{i \in s}(x_i - \hat{\bar{x}})^2}$$

şi

$$\hat{\alpha} = \hat{\bar{y}} - \hat{\beta} \cdot \hat{\bar{x}}$$

La nivel de eşantion, estimaţia parametrului medie obţinută cu această metodă se determină pe baza relaţiei:

$$\bar{y}_{reg} = \bar{y} + b(\bar{X} - \bar{x})$$

Relaţia se poate scrie şi în funcţie de coeficientul de corelaţie dintre cele două variabile, ştiind că:

$$b = r_{xy}\frac{s_x}{s_y}$$

Rezultă:

$$\bar{y}_{reg} = \bar{y} + r_{xy}\frac{s_x}{s_y}(\bar{X} - \bar{x})$$

Observaţie

Ideea de bază a acestui estimator este să se estimeze media populaţiei totale ($\bar{Y}$) prin intermediul punctului de pe dreapta de regresie care are abscisa $\bar{X}$, punct obţinut prin metoda celor mai mici pătrate.

Pentru $\alpha = 0$, problema se reduce la cazul estimatorului raport. În plus, se observă că β este estimat cu ajutorul estimatorului $\hat{\beta}$, care este un estimator de tip raport.

5.4.2. Proprietăţile estimatorului

a. Nedeplasarea

Estimatorul regresie este un estimator uşor deplasat, cu o deplasare care are un ordin de mărime în $1/n$ egal cu unu. Pentru eşantioane de volum mare, această deplasare devine neglijabilă, însă pentru eşantioane de volum mic ($n < 30$), deplasarea este semnificativă.

b. Varianța

Ca și în cazul estimatorului raport, varianța estimatorului regresie este:

$$V\left(\hat{\bar{Y}}_{reg}\right) \cong \frac{1}{n}\sigma_u^2$$

unde

$$u_i = y_i - \hat{\alpha} - \hat{\beta} \cdot x_i$$

și

$$\sigma_u^2 = \frac{1}{N-1}\sum_{i=1}^{N} u_i^2$$

Dar

$$\sigma_u^2 = \sigma_Y^2 + \hat{\beta}^2\sigma_X^2 - 2\hat{\beta}\sigma_{XY} = \sigma_Y^2 + \left(\frac{\sigma_{XY}}{\sigma_X^2}\right)^2 \sigma_X^2 - 2\frac{\sigma_{XY}}{\sigma_X^2}\sigma_{XY}$$

De unde rezultă că:

$$\sigma_u^2 = \sigma_Y^2 - \left(\frac{\sigma_{XY}}{\sigma_X}\right)^2 = \sigma_Y^2 - \rho(X,Y)\sigma_Y^2 = \sigma_Y^2\big(1 - \rho(X,Y)\big)$$

În concluzie,

$$V\left(\hat{\bar{Y}}_{reg}\right) \cong \frac{1}{n}\sigma_u^2 = \frac{1}{n}\sigma_Y^2\big(1 - \rho(X,Y)\big) = V(\hat{\bar{y}})\big(1 - \rho(X,Y)\big)$$

Varianța estimatorului regresie este dată prin relația:

$$V\left(\hat{\bar{Y}}_{reg}\right) = V(\hat{\bar{y}})\big(1 - \rho(X,Y)\big)$$

relație care arată că prin acest estimator s-a obținut un plus de precizie. Varianța estimatorului regresie este întotdeauna mai mică decât a estimatorului inițial, fiind egale doar în cazul în care între variabilele X și Y nu există nici o legătură (coeficientul de corelație este zero).

Se recomandă ca această metodă de ameliorare să fie folosită în toate cazurile în care se dispune de o variabilă X cunoscută la nivelul populației de referință și care este corelată cu variabila de interes Y.

Pentru calculul unui interval de încredere, se poate estima varianța estimatorului raport cu ajutorul relației:

$$s^2_{\hat{\bar{Y}}_{reg}} = s_Y'^2\big(1 - r_{xy}\big)$$

unde r_{xy} este estimația coeficientului de corelație simplă pentru cele două variabile, calculată la nivelul eșantionului.

Observație

Se poate observa că estimatorul regresie este un caz particular al estimatorului de forma:

$$\hat{\bar{Y}}_{dif} = \hat{\bar{y}} + \gamma(\bar{X} - \hat{\bar{x}})$$

numit *estimator prin diferență.*

Dacă este cunoscută valoarea lui γ ($\gamma = \gamma_0$), estimatorul se numește *estimator prin diferență generalizat*, iar dacă valoarea lui γ este necunoscută, suntem în cazul estimatorului regresie.

Estimatorul prin diferență generalizat are următoarele proprietăți:

- este nedeplasat:

$$E\left(\hat{\bar{Y}}_{dif}\right) = E(\hat{\bar{y}}) + \gamma_0 \cdot E(\bar{X} - \hat{\bar{x}}) = \bar{Y}$$

- varianța estimatorului este dată de relația:

$$V\left(\hat{\bar{Y}}_{dif}\right) = V(\hat{\bar{y}}) + \gamma_0^2 \cdot V(\hat{\bar{x}}) - 2\gamma_0 cov(\hat{\bar{x}}, \hat{\bar{y}})$$

$$V\left(\hat{\bar{Y}}_{dif}\right) = \frac{1}{n}(\sigma_Y^2 + \gamma_0^2\sigma_X^2 - 2\gamma_0\rho(X,Y)\sigma_X\sigma_Y)$$

Estimatorul prin diferență generalizat aduce o îmbunătățire a preciziei dacă se poate obține acea valoare a lui γ_0 care să determine o varianță minimă a diferențelor:

$$d_i = y_i - \gamma_0 x_i$$

Valoarea lui γ_0 se poate obține din cercetări anterioare sau prin ședințe precum brainstorming.

Exemplu

Considerăm un sondaj aleator simplu realizat cu obiectivul de a estima producția medie de cereale boabe la nivelul unei unități agricole dintr-un județ. În acest scop, s-a constituit un eșantion reprezentativ format din cinci unități agricole ale județului, pentru care s-au obținut rezultatele din tabelul 5.3 (date convenționale).

Tabelul 5.3. *Producția de cereale și suprafața cultivată*

Unitate	Producția *Y* (tone)	Suprafața *X* (ha)
1	150	79
2	169	96
3	260	107
4	292	148
5	145	76
Total	**1016**	**506**

Cunoscând aceste rezultate şi că suprafaţa medie cultivată cu cereale într-o unitate agricolă a judeţului este de 110 ha, se poate redresa calitatea estimatorului producţiei medii de cereale boabe cu ajutorul estimatorului regresie.

Estimatorul clasic al unui total este

$$\hat{\bar{y}} = \frac{1}{n}\sum_{i \in s} y_i$$

iar la nivelul eşantionului se calculează estimaţia:

$$\bar{y} = \frac{1}{5}\sum_{i=1}^{5} y_i = 203{,}2 \text{ t}$$

Estimatorul clasic poate fi îmbunătăţit considerând informaţia suplimentară oferită de variabila X - suprafaţa cultivată.

Estimatorul regresie are expresia:

$$\hat{\bar{Y}}_{reg} = \hat{\bar{y}} + \hat{\beta}(\bar{X} - \hat{\bar{x}})$$

unde

$$\hat{\beta} = \frac{\sum_{i \in s}(x_i - \hat{\bar{x}})(y_i - \hat{\bar{y}})}{\sum_{i \in s}(x_i - \hat{\bar{x}})^2}$$

La nivelul eşantionului extras, estimaţia parametrului de regresie se obţine prin relaţia:

$$b = \frac{5\sum_{i=1}^{5} x_i \cdot y_i - \sum_{i=1}^{5} x_i \cdot \sum_{i=1}^{5} y_i}{5\sum_{i=1}^{5} x_i^2 - \left(\sum_{i=1}^{5} x_i\right)^2}$$

Conform datelor din tabelul 5.4, se poate obţine estimaţia parametrului de regresie:

$$b = \frac{5 \cdot 110130 - 1016 \cdot 506}{5 \cdot 54586 - 256036} = 2{,}16$$

Tabelul 5.4. *Elemente de calcul*

	y_i	x_i	$y_i x_i$	x_i^2
1	150	79	11850	6241
2	169	96	16224	9216
3	260	107	27820	11449
4	292	148	43216	21904
5	145	76	11020	5776
Total	**1016**	**506**	**110130**	**54586**

Înlocuind în expresia estimatorului, se obţine estimaţia parametrului medie cu ajutorul estimatorului regresie:

$$\bar{y}_{reg} = \bar{y} + b(\bar{X} - \bar{x}) = 203{,}2 + 2{,}16(110 - 101{,}2) = 222{,}2 \text{ t}$$

Se observă că estimația obținută în urma redresării este mai mare decât cea obținută clasic, însă important este că această valoare este preferabilă celei obținute cu estimatorul clasic în sensul preciziei mai ridicate.

5.5. Exerciții

1. Se realizează o anchetă asupra studenților (un eșantion de 100 de persoane) unei universități (care are un total de 5000 de studenți), pentru a testa abilitățile de comunicare în scris cu ajutorul unui test. Rezultatele testului (ponderea celor care trec testul) și structura studenților universității după profilul liceului absolvit (Real, Uman, Vocațional) sunt prezentate în tabelul 5.5.

Tabelul 5.5. *Rezultate anchetă și structură populație*

Profilul	$N_i/N \times 100$	$n_i/n \times 100$
Real	45	15
Uman	35	40
Vocațional	20	25
Total	**100**	**80**

Se cere:

- să se estimeze punctual și prin interval de încredere procentul studenților care trec examenul la nivelul întregii universități;
- să se estimeze procentul celor care trec testul utilizând informația suplimentară privind structura studenților după liceul absolvit.

2. În ancheta prezentată la exercițiul 1, s-au înregistrat datele și s-au calculat estimațiile pentru punctajul obținut (variabila X) de studenții din eșantion. Rezultatele sunt prezentate în tabelul 5.6.

Tabelul 5.6. *Rezultate anchetă și structură populație*

Profilul	$s_h'^2$	$\bar{x}_h$	n_h	N_h
Real	100	70	15	2250
Uman	120	85	40	1750
Vocațional	90	80	25	1000
Total	-	-	**100**	**5000**

Se cere:

- să se estimeze punctual și prin interval de încredere punctajul mediu al studenților la nivelul întregii universități;

- cu ajutorul informației privind structura populației studenților, să se estimeze punctual și prin interval de încredere punctajul mediu la nivelul universității.

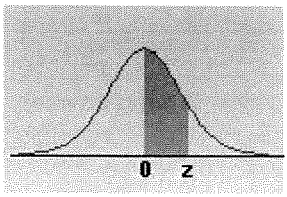
3. Se consideră o anchetă (un eșantion de 100 de persoane) la nivelul studenților unei facultăți privind capacitatea de analiză (variabila de interes Y) a unei anumite probleme complexe. Rezultatele unui test au fost cuantificate cu un punctaj, iar la nivel de eșantion s-a obținut un punctaj mediu de 8,5 și o abatere standard de 1,2 puncte. De asemenea, la nivel de eșantion se cunosc rezultatele la examenul de matematică (variabila suplimentară X), cu o medie de 7,9 puncte și o abatere standard de 2 puncte. Asumând ipoteza că există o legătură semnificativă între cele două variabile (coeficientul de corelație simplă estimat pentru cele două variabile este egal cu 0,82), se cere:

- să se estimeze punctual și prin interval de încredere punctajul mediu pentru variabila de interes utilizând un estimator de tip raport;
- se cere același lucru ca mai sus, în condițiile utilizării unui estimator de tip regresie. Se cunoaște că media la examenul de matematică, la nivelul populației totale este egală cu 8,2 puncte;
- să se compare cele două rezultate.

Tabele probabiliste

Funcţia Laplace

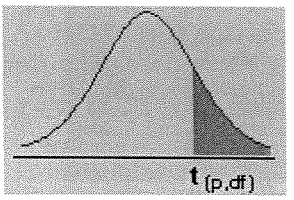
$$\Phi(z) = \int_0^z e^{-\frac{t^2}{2}} dt$$



	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Repartiția Student

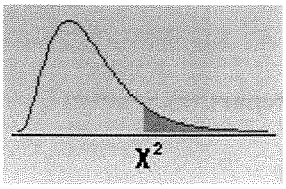
$p = P(t > t_{p,n})$



n\p	0,10	0,05	0,025	0,01	0,005
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
n>30	1,282	1,645	1,960	2,326	2,576

Repartiţia Chi-pătrat

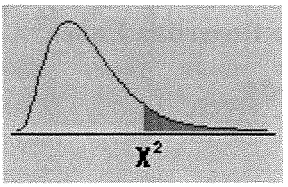
$$p = P(\chi^2 > \chi^2_{p,n})$$



n\p	0,100	0,050	0,025	0,010	0,005
1	2,706	3,841	5,024	6,635	7,879
2	4,605	5,991	7,378	9,210	10,597
3	6,251	7,815	9,348	11,345	12,838
4	7,779	9,488	11,143	13,277	14,860
5	9,236	11,071	12,833	15,086	16,750
6	10,645	12,592	14,449	16,812	18,548
7	12,017	14,067	16,013	18,475	20,278
8	13,362	15,507	17,535	20,090	21,955
9	14,684	16,919	19,023	21,666	23,589
10	15,987	18,307	20,483	23,209	25,188
11	17,275	19,675	21,920	24,725	26,757
12	18,549	21,026	23,337	26,217	28,300
13	19,812	22,362	24,736	27,688	29,819
14	21,064	23,685	26,119	29,141	31,319
15	22,307	24,996	27,488	30,578	32,801
16	23,542	26,296	28,845	32,000	34,267
17	24,769	27,587	30,191	33,409	35,718
18	25,989	28,869	31,526	34,805	37,156
19	27,204	30,144	32,852	36,191	38,582
20	28,412	31,410	34,170	37,566	39,997
21	29,615	32,671	35,479	38,932	41,401
22	30,813	33,924	36,781	40,289	42,796
23	32,007	35,172	38,076	41,638	44,181
24	33,196	36,415	39,364	42,980	45,559
25	34,382	37,652	40,646	44,314	46,928
26	35,563	38,885	41,923	45,642	48,290
27	36,741	40,113	43,195	46,963	49,645
28	37,916	41,337	44,461	48,278	50,993
29	39,087	42,557	45,722	49,588	52,336
30	40,256	43,773	46,979	50,892	53,672

Repartiţia Chi-pătrat

$$p = P(\chi^2 > \chi^2_{p,n})$$

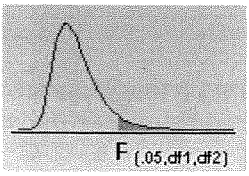


n\p	0,995	0,975	0,950	0,990	0,900
1	0,00004	0,00016	0,00098	0,00393	0,01579
2	0,01003	0,02010	0,05064	0,10259	0,21072
3	0,07172	0,11483	0,21580	0,35185	0,58437
4	0,20699	0,29711	0,48442	0,71072	1,06362
5	0,41174	0,55430	0,83121	1,14548	1,61031
6	0,67573	0,87209	1,23734	1,63538	2,20413
7	0,98926	1,23904	1,68987	2,16735	2,83311
8	1,34441	1,64650	2,17973	2,73264	3,48954
9	1,73493	2,08790	2,70039	3,32511	4,16816
10	2,15586	2,55821	3,24697	3,94030	4,86518
11	2,60322	3,05348	3,81575	4,57481	5,57778
12	3,07382	3,57057	4,40379	5,22603	6,30380
13	3,56503	4,10692	5,00875	5,89186	7,04150
14	4,07467	4,66043	5,62873	6,57063	7,78953
15	4,60092	5,22935	6,26214	7,26094	8,54676
16	5,14221	5,81221	6,90766	7,96165	9,31224
17	5,69722	6,40776	7,56419	8,67176	10,08519
18	6,26480	7,01491	8,23075	9,39046	10,86494
19	6,84397	7,63273	8,90652	10,11701	11,65091
20	7,43384	8,26040	9,59078	10,85081	12,44261
21	8,03365	8,89720	10,28290	11,59131	13,23960
22	8,64272	9,54249	10,98232	12,33801	14,04149
23	9,26042	10,19572	11,68855	13,09051	14,84796
24	9,88623	10,85636	12,40115	13,84843	15,65868
25	10,51965	11,52398	13,11972	14,61141	16,47341
26	11,16024	12,19815	13,84390	15,37916	17,29188
27	11,80759	12,87850	14,57338	16,15140	18,11390
28	12,46134	13,56471	15,30786	16,92788	18,93924
29	13,12115	14,25645	16,04707	17,70837	19,76774
30	13,78672	14,95346	16,79077	18,49266	20,59923

Repartiţia Fisher

$\alpha = 0,05$

$df_1 = n_1$, $df_2 = n_2$



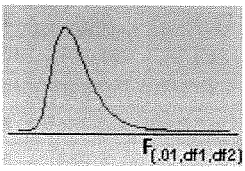
n_2/n_1	1	2	3	4	5	6	7
1	161,448	199,500	215,707	224,583	230,162	233,986	236,768
2	18,513	19,000	19,164	19,247	19,296	19,330	19,353
3	10,128	9,552	9,277	9,117	9,014	8,941	8,887
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094
5	6,608	5,786	5,410	5,192	5,050	4,950	4,876
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787
8	5,318	4,459	4,066	3,838	3,688	3,581	3,501
9	5,117	4,257	3,863	3,633	3,482	3,374	3,293
10	4,965	4,103	3,708	3,478	3,326	3,217	3,136
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764
15	4,543	3,682	3,287	3,056	2,901	2,791	2,707
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514
21	4,325	3,467	3,073	2,840	2,685	2,573	2,488
22	4,301	3,443	3,049	2,817	2,661	2,549	2,464
23	4,279	3,422	3,028	2,796	2,640	2,528	2,442
24	4,260	3,403	3,009	2,776	2,621	2,508	2,423
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405
26	4,225	3,369	2,975	2,743	2,587	2,474	2,388
27	4,210	3,354	2,960	2,728	2,572	2,459	2,373
28	4,196	3,340	2,947	2,714	2,558	2,445	2,359
29	4,183	3,328	2,934	2,701	2,545	2,432	2,346
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334
40	4,085	3,232	2,839	2,606	2,450	2,336	2,249
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087
n>120	3,842	2,996	2,605	2,372	2,214	2,099	2,010

n_2/n_1	8	9	10	20	30	120	$n_1>120$
1	238,883	240,543	241,882	248,013	250,095	253,253	254,314
2	19,371	19,385	19,396	19,446	19,462	19,487	19,496
3	8,845	8,812	8,786	8,660	8,617	8,549	8,526
4	6,041	5,999	5,964	5,803	5,746	5,658	5,628
5	4,818	4,773	4,735	4,558	4,496	4,399	4,365
6	4,147	4,099	4,060	3,874	3,808	3,705	3,669
7	3,726	3,677	3,637	3,445	3,376	3,267	3,230
8	3,438	3,388	3,347	3,150	3,079	2,967	2,928
9	3,230	3,179	3,137	2,937	2,864	2,748	2,707
10	3,072	3,020	2,978	2,774	2,700	2,580	2,538
11	2,948	2,896	2,854	2,646	2,571	2,448	2,405
12	2,849	2,796	2,753	2,544	2,466	2,341	2,296
13	2,767	2,714	2,671	2,459	2,380	2,252	2,206
14	2,699	2,646	2,602	2,388	2,308	2,178	2,131
15	2,641	2,588	2,544	2,328	2,247	2,114	2,066
16	2,591	2,538	2,494	2,276	2,194	2,059	2,010
17	2,548	2,494	2,450	2,230	2,148	2,011	1,960
18	2,510	2,456	2,412	2,191	2,107	1,968	1,917
19	2,477	2,423	2,378	2,156	2,071	1,930	1,878
20	2,447	2,393	2,348	2,124	2,039	1,896	1,843
21	2,421	2,366	2,321	2,096	2,010	1,866	1,812
22	2,397	2,342	2,297	2,071	1,984	1,838	1,783
23	2,375	2,320	2,275	2,048	1,961	1,813	1,757
24	2,355	2,300	2,255	2,027	1,939	1,790	1,733
25	2,337	2,282	2,237	2,008	1,919	1,768	1,711
26	2,321	2,266	2,220	1,990	1,901	1,749	1,691
27	2,305	2,250	2,204	1,974	1,884	1,731	1,672
28	2,291	2,236	2,190	1,959	1,869	1,714	1,654
29	2,278	2,223	2,177	1,945	1,854	1,698	1,638
30	2,266	2,211	2,165	1,932	1,841	1,684	1,622
40	2,180	2,124	2,077	1,839	1,744	1,577	1,509
60	2,097	2,040	1,993	1,748	1,649	1,467	1,389
120	2,016	1,959	1,911	1,659	1,554	1,352	1,254
$n_2>120$	1,938	1,880	1,831	1,571	1,459	1,221	1,000

Repartiția Fisher

$\alpha = 0,01$

$df_1 = n_1$, $df_2 = n_2$



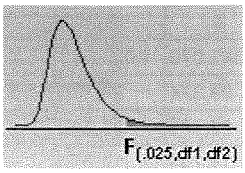
n_2/n_1	1	2	3	4	5	6	7
1	4052,181	4999,500	5403,352	5624,583	5763,650	5858,986	5928,356
2	98,503	99,000	99,166	99,249	99,299	99,333	99,356
3	34,116	30,817	29,457	28,710	28,237	27,911	27,672
4	21,198	18,000	16,694	15,977	15,522	15,207	14,976
5	16,258	13,274	12,060	11,392	10,967	10,672	10,456
6	13,745	10,925	9,780	9,148	8,746	8,466	8,260
7	12,246	9,547	8,451	7,847	7,460	7,191	6,993
8	11,259	8,649	7,591	7,006	6,632	6,371	6,178
9	10,561	8,022	6,992	6,422	6,057	5,802	5,613
10	10,044	7,559	6,552	5,994	5,636	5,386	5,200
11	9,646	7,206	6,217	5,668	5,316	5,069	4,886
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640
13	9,074	6,701	5,739	5,205	4,862	4,620	4,441
14	8,862	6,515	5,564	5,035	4,695	4,456	4,278
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142
16	8,531	6,226	5,292	4,773	4,437	4,202	4,026
17	8,400	6,112	5,185	4,669	4,336	4,102	3,927
18	8,285	6,013	5,092	4,579	4,248	4,015	3,841
19	8,185	5,926	5,010	4,500	4,171	3,939	3,765
20	8,096	5,849	4,938	4,431	4,103	3,871	3,699
21	8,017	5,780	4,874	4,369	4,042	3,812	3,640
22	7,945	5,719	4,817	4,313	3,988	3,758	3,587
23	7,881	5,664	4,765	4,264	3,939	3,710	3,539
24	7,823	5,614	4,718	4,218	3,895	3,667	3,496
25	7,770	5,568	4,675	4,177	3,855	3,627	3,457
26	7,721	5,526	4,637	4,140	3,818	3,591	3,421
27	7,677	5,488	4,601	4,106	3,785	3,558	3,388
28	7,636	5,453	4,568	4,074	3,754	3,528	3,358
29	7,598	5,420	4,538	4,045	3,725	3,499	3,330
30	7,562	5,390	4,510	4,018	3,699	3,473	3,304
40	7,314	5,179	4,313	3,828	3,514	3,291	3,124
60	7,077	4,977	4,126	3,649	3,339	3,119	2,953
120	6,851	4,787	3,949	3,480	3,174	2,956	2,792
n_2>120	6,635	4,605	3,782	3,319	3,017	2,802	2,639

n_2/n_1	8	9	10	20	30	120	$n_1>120$
1	5981,070	6022,473	6055,847	6208,730	6260,649	6339,391	6365,864
2	99,374	99,388	99,399	99,449	99,466	99,491	99,499
3	27,489	27,345	27,229	26,690	26,505	26,221	26,125
4	14,799	14,659	14,546	14,020	13,838	13,558	13,463
5	10,289	10,158	10,051	9,553	9,379	9,112	9,020
6	8,102	7,976	7,874	7,396	7,229	6,969	6,880
7	6,840	6,719	6,620	6,155	5,992	5,737	5,650
8	6,029	5,911	5,814	5,359	5,198	4,946	4,859
9	5,467	5,351	5,257	4,808	4,649	4,398	4,311
10	5,057	4,942	4,849	4,405	4,247	3,996	3,909
11	4,744	4,632	4,539	4,099	3,941	3,690	3,602
12	4,499	4,388	4,296	3,858	3,701	3,449	3,361
13	4,302	4,191	4,100	3,665	3,507	3,255	3,165
14	4,140	4,030	3,939	3,505	3,348	3,094	3,004
15	4,004	3,895	3,805	3,372	3,214	2,959	2,868
16	3,890	3,780	3,691	3,259	3,101	2,845	2,753
17	3,791	3,682	3,593	3,162	3,003	2,746	2,653
18	3,705	3,597	3,508	3,077	2,919	2,660	2,566
19	3,631	3,523	3,434	3,003	2,844	2,584	2,489
20	3,564	3,457	3,368	2,938	2,778	2,517	2,421
21	3,506	3,398	3,310	2,880	2,720	2,457	2,360
22	3,453	3,346	3,258	2,827	2,667	2,403	2,305
23	3,406	3,299	3,211	2,781	2,620	2,354	2,256
24	3,363	3,256	3,168	2,738	2,577	2,310	2,211
25	3,324	3,217	3,129	2,699	2,538	2,270	2,169
26	3,288	3,182	3,094	2,664	2,503	2,233	2,131
27	3,256	3,149	3,062	2,632	2,470	2,198	2,097
28	3,226	3,120	3,032	2,602	2,440	2,167	2,064
29	3,198	3,092	3,005	2,574	2,412	2,138	2,034
30	3,173	3,067	2,979	2,549	2,386	2,111	2,006
40	2,993	2,888	2,801	2,369	2,203	1,917	1,805
60	2,823	2,718	2,632	2,198	2,028	1,726	1,601
120	2,663	2,559	2,472	2,035	1,860	1,533	1,381
$n_2>120$	2,511	2,407	2,321	1,878	1,696	1,325	1,000

Repartiția Fisher

$\alpha = 0,025$

$df_1 = n_1$, $df_2 = n_2$



n_2/n_1	1	2	3	4	5	6	7
1	647,789	799,500	864,163	899,583	921,847	937,111	948,216
2	385,063	390,000	391,655	392,484	392,982	393,315	393,552
3	174,434	160,441	154,392	151,010	148,848	147,347	146,244
4	122,179	106,491	99,792	96,045	93,645	91,973	90,741
5	100,070	84,336	77,636	73,879	71,464	69,777	68,531
6	88,131	72,599	65,988	62,272	59,876	58,198	56,955
7	80,727	65,415	58,898	55,226	52,852	51,186	49,949
8	75,709	60,595	54,160	50,526	48,173	46,517	45,286
9	72,093	57,147	50,781	47,181	44,844	43,197	41,970
10	69,367	54,564	48,256	44,683	42,361	40,721	39,498
11	67,241	52,559	46,300	42,751	40,440	38,807	37,586
12	65,538	50,959	44,742	41,212	38,911	37,283	36,065
13	64,143	49,653	43,472	39,959	37,667	36,043	34,827
14	62,979	48,567	42,417	38,919	36,634	35,014	33,799
15	61,995	47,650	41,528	38,043	35,764	34,147	32,934
16	61,151	46,867	40,768	37,294	35,021	33,406	32,194
17	60,420	46,189	40,112	36,648	34,379	32,767	31,556
18	59,781	45,597	39,539	36,083	33,820	32,209	30,999
19	59,216	45,075	39,034	35,587	33,327	31,718	30,509
20	58,715	44,613	38,587	35,147	32,891	31,283	30,074
21	58,266	44,199	38,188	34,754	32,501	30,895	29,686
22	57,863	43,828	37,829	34,401	32,151	30,546	29,338
23	57,498	43,492	37,505	34,083	31,835	30,232	29,023
24	57,166	43,187	37,211	33,794	31,548	29,946	28,738
25	56,864	42,909	36,943	33,530	31,287	29,685	28,478
26	56,586	42,655	36,697	33,289	31,048	29,447	28,240
27	56,331	42,421	36,472	33,067	30,828	29,228	28,021
28	56,096	42,205	36,264	32,863	30,626	29,027	27,820
29	55,878	42,006	36,072	32,674	30,438	28,840	27,633
30	55,675	41,821	35,894	32,499	30,265	28,667	27,460
40	54,239	40,510	34,633	31,261	29,037	27,444	26,238
60	52,856	39,253	33,425	30,077	27,863	26,274	25,068
120	51,523	38,046	32,269	28,943	26,740	25,154	23,948
n>120	50,239	36,889	31,161	27,858	25,665	24,082	22,875

n_2/n_1	8	9	10	20	30	120	$n_1>120$
1	956,656	963,284	968,627	993,102	1001,414	1014,020	1018,258
2	393,730	393,869	393,980	394,479	39,465	39,490	39,498
3	145,399	144,731	144,189	141,674	14,081	13,947	13,902
4	89,796	89,047	88,439	85,599	8,461	8,309	8,257
5	67,572	66,811	66,192	63,286	6,227	6,069	6,015
6	55,996	55,234	54,613	51,684	5,065	4,904	4,849
7	48,993	48,232	47,611	44,667	4,362	4,199	4,142
8	44,333	43,572	42,951	39,995	3,894	3,728	3,670
9	41,020	40,260	39,639	36,669	3,560	3,392	3,333
10	38,549	37,790	37,168	34,185	3,311	3,140	3,080
11	36,638	35,879	35,257	32,261	3,118	2,944	2,883
12	35,118	34,358	33,736	30,728	2,963	2,787	2,725
13	33,880	33,120	32,497	29,477	2,837	2,659	2,595
14	32,853	32,093	31,469	28,437	2,732	2,552	2,487
15	31,987	31,227	30,602	27,559	2,644	2,461	2,395
16	31,248	30,488	29,862	26,808	2,568	2,383	2,316
17	30,610	29,849	29,222	26,158	2,502	2,315	2,247
18	30,053	29,291	28,664	25,590	2,445	2,256	2,187
19	29,563	28,801	28,172	25,089	2,394	2,203	2,133
20	29,128	28,365	27,737	24,645	2,349	2,156	2,085
21	28,740	27,977	27,348	24,247	2,308	2,114	2,042
22	28,392	27,628	26,998	23,890	2,272	2,076	2,003
23	28,077	27,313	26,682	23,567	2,239	2,041	1,968
24	27,791	27,027	26,396	23,273	2,209	2,010	1,935
25	27,531	26,766	26,135	23,005	2,182	1,981	1,906
26	27,293	26,528	25,896	22,759	2,157	1,954	1,878
27	27,074	26,309	25,676	22,533	2,133	1,930	1,853
28	26,872	26,106	25,473	22,324	2,112	1,907	1,829
29	26,686	25,919	25,286	22,131	2,092	1,886	1,807
30	26,513	25,746	25,112	21,952	2,074	1,866	1,787
40	25,289	24,519	23,882	20,677	1,943	1,724	1,637
60	24,117	23,344	22,702	19,445	1,815	1,581	1,482
120	22,994	22,217	21,570	18,249	1,690	1,433	1,310
$n_2>120$	21,918	21,136	20,483	17,085	1,566	1,268	1,000

Repartiţia Durbin-Watson

$\alpha = 0,05$; k reprezintă numărul de parametri din model

	k=2		*k=3*		*k=4*		*k=5*	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
7	0,700	1,356	0,467	1,896	-----	-----	-----	-----
8	0,763	1,332	0,559	1,777	0,367	2,287	-----	-----
9	0,824	1,320	0,629	1,699	0,455	2,128	0,296	2,588
10	0,879	1,320	0,697	1,641	0,525	2,016	0,376	2,414
11	0,927	1,324	0,758	1,604	0,595	1,928	0,444	2,283
12	0,971	1,331	0,812	1,579	0,658	1,864	0,512	2,177
13	1,010	1,340	0,861	1,562	0,715	1,816	0,574	2,094
14	1,045	1,350	0,905	1,551	0,767	1,779	0,632	2,030
15	1,077	1,361	0,946	1,543	0,814	1,750	0,685	1,977
16	1,106	1,371	0,982	1,539	0,857	1,728	0,734	1,935
17	1,133	1,381	1,015	1,536	0,897	1,710	0,779	1,900
18	1,158	1,391	1,046	1,535	0,933	1,696	0,820	1,872
19	1,180	1,401	1,074	1,536	0,967	1,685	0,859	1,848
20	1,201	1,411	1,100	1,537	0,998	1,676	0,894	1,828
21	1,221	1,420	1,125	1,538	1,026	1,669	0,927	1,812
22	1,239	1,429	1,147	1,541	1,053	1,664	0,958	1,797
23	1,257	1,437	1,168	1,543	1,078	1,660	0,986	1,785
24	1,273	1,446	1,188	1,546	1,101	1,656	1,013	1,775
25	1,288	1,454	1,206	1,550	1,123	1,654	1,038	1,767
26	1,302	1,461	1,224	1,553	1,143	1,652	1,062	1,759
27	1,316	1,469	1,240	1,556	1,162	1,651	1,084	1,753
28	1,328	1,476	1,255	1,560	1,181	1,650	1,104	1,747
29	1,341	1,483	1,270	1,563	1,198	1,650	1,124	1,743
30	1,352	1,489	1,284	1,567	1,214	1,650	1,143	1,739
31	1,363	1,496	1,297	1,570	1,229	1,650	1,160	1,735
32	1,373	1,502	1,309	1,574	1,244	1,650	1,177	1,732
33	1,383	1,508	1,321	1,577	1,258	1,651	1,193	1,730
34	1,393	1,514	1,333	1,580	1,271	1,652	1,208	1,728
35	1,402	1,519	1,343	1,584	1,283	1,653	1,222	1,726
36	1,411	1,525	1,354	1,587	1,295	1,654	1,236	1,724
37	1,419	1,530	1,364	1,590	1,307	1,655	1,249	1,723
38	1,427	1,535	1,373	1,594	1,318	1,656	1,261	1,722
39	1,435	1,540	1,382	1,597	1,328	1,658	1,273	1,722
40	1,442	1,544	1,391	1,600	1,338	1,659	1,285	1,721
45	1,475	1,566	1,430	1,615	1,383	1,666	1,336	1,720
50	1,503	1,585	1,462	1,628	1,421	1,674	1,378	1,721
55	1,528	1,601	1,490	1,641	1,452	1,681	1,414	1,724
60	1,549	1,616	1,514	1,652	1,480	1,689	1,444	1,727
65	1,567	1,629	1,536	1,662	1,503	1,696	1,471	1,731
70	1,583	1,641	1,554	1,672	1,525	1,703	1,494	1,735
75	1,598	1,652	1,571	1,680	1,543	1,709	1,515	1,739
80	1,611	1,662	1,586	1,688	1,560	1,715	1,534	1,743
85	1,624	1,671	1,600	1,696	1,575	1,721	1,550	1,747
90	1,635	1,679	1,612	1,703	1,589	1,726	1,566	1,751
95	1,645	1,687	1,623	1,709	1,602	1,732	1,579	1,755
100	1,654	1,694	1,634	1,715	1,613	1,736	1,592	1,758
150	1,720	1,747	1,706	1,760	1,693	1,774	1,679	1,788
200	1,758	1,779	1,748	1,789	1,738	1,799	1,728	1,809

Bibliografie

Antoine, J. (1969), *L'opinion. Techniques d'enquêtes par sondage*, Éditions Dunod, Paris.

Antoine, J. (1990), *Le sondage outil du marketing*, Éditions Dunod, Paris.

Ardilly, P. (1994), *Les techniques de sondage*, Éditions Technip, Paris.

Ardilly, P., Tillé, Y. (2006), *Sampling methods. Exercises and solutions*, Springer-Verlag, New York.

Arnab, R. (1991), On sampling over two occasions using varying probabilities, *Journal of the Indian Society of Agricultural Statistics*, 43, pp. 282-290.

Arnab, R. (2017), *Survey sampling theory and applications*, Academic Press, London.

Bărbat, A. (1971), *Teoria statisticii sociale*, Editura Universității „Alexandru Ioan Cuza" din Iași.

Bechhofer, F., Paterson, L. (2000), *Principles of research design in the social science*, Routledge, London.

Berrebi, L. (1992), Estimation par régression simple. *Méthodes*, INSEE, (29-30-31), pp. 239-264.

Berthier, N. (2002), *Les techniques d'enquête*, Éditions Armand Colin, Paris.

Bouget, D., Vienot, A. (1995), *Traitement de l'information statistiques et probabilités*, Éditions Vuibert, Paris.

Chardon, P.A. (1981), *Méthodes pratiques de dépouillement de questionnaires*, Thèse, Université de Neuchâtel, Imprimerie de l'Ouest SA, Peseux.

Chaudhuri, A., Stenger, H. (2005), *Survey sampling: theory and methods*, CRC Press, London.

Chelcea, S. (1975), *Chestionarul în investigația sociologică*, Editura Științifică și Enciclopedică, București.

Chelcea, S. (2004), *Inițiere în cercetarea sociologică*, Editura Comunicare, București.

Cochran, W.G. (1977), *Sampling techniques*, John Wiley & Sons, London.

Cozby, P.C. (2011), *Methods in behavioral research*, McGraw-Hill Education.

Daudin, J., Robin, S. (1999), *Statistique inférentielle*, Presses Universitaires de Rennes.

Desabie, J. (1966), *Théorie et pratique des sondages*, Éditions Dunod, Paris.

Desrosiere, A. (2000), *La politique des grands nombre*, Éditions La Découverte, Paris.

Droesbeke, J.J., Fichet, B., Tassi, P. (1987), *Les sondages*, Éditions Economica, Paris.

Droesbeke, J.J., Tassi, P. (1990), *Histoire de la statistique*, Presses Universitaires de France, Paris.

Dumas de Rauly, D. (1966), *L'estimation statistique*, Éditions Gauthier-Villars, Paris.
Durkheim, E. (1924), *Regulile metodei sociologice*, Editura Cultura Națională, București.
Dussaix, A.M., Grosbras, J.M. (1992), *Exercices de sondages*, Éditions Economica, Paris.
Dussaix, A.M., Brossier, G. (1999), *Enquêtes et sondages. Méthodes, modèles, applications, nouvelles approches*, Éditions Dunod, Paris.
Dussaix, A.M. (1993), *Les sondages: principes et méthodes*, Presses Universitaires de France, Paris.
Ericson, W.A. (1969b), A note on the posterior mean of a population mean, *Journal of the Royal Statistical Society*, Series B, 31, pp. 332-334.
Fisher, R.A. (1947), *Les méthodes statistiques*, Presses Universitaires de France, Paris.
Fink, A. (1995), *How to report on surveys*, Sage Publications, London.
Fuller, W.A. (2009), *Sampling statistics*, John Wiley & Sons, Hoboken, New Jersey.
Ghiglione, R., Matalon, R. (1985), *Les enquêtes sociologiques. Théories et pratique*, Éditions Armand Colin, Paris.
Gillham, B. (2002), *Developing a questionnaire*, Continuum, London.
Goode, W., Hatt, P. (1952), *Methods in social research*, McGraw-Hill, New York.
Gourieroux, G. (1981), *Théorie des sondages*, Éditions Economica, Paris.
Grais, B. (1998), *Méthodes statistiques*, Éditions Dunod, Paris.
Grange, D., Lebart, L. (1994), *Traitements statistiques des enquêtes*, Éditions Dunod, Paris.
Duverger, M. (1971), *Méthodes des sciences sociales*, Presses Universitaires de Paris.
Gravitz, M. (2001), *Méthodes des sciences sociales*, Éditions Dalloz, Paris.
Grosbras, J.M. (1987), *Méthodes statistiques des sondages*, Éditions Economica, Paris.
Holt, D., Smith, T.M. (1979), Post stratification, *Journal of the Royal Statistical Society*, 142(A), pp. 33-46.
Hugues, J. (1988), *Qui? Quoi? Comment? Ou la pratique des sondages*, Eyrolles, Paris.
Jaba, E. (2002), *Statistica*, Ediția a treia, Editura Economică, București.
Javeau, Cl. (1978), *L'enquête par questionnaire*, Editions de l'Université de Bruxelles.
Jemna, D.V. (2004), O structură de gândire asupra chestionarului statistic, *Probleme actuale de statistică*, Junimea, Iași, 2004, pp. 113-121.
Jemna, D.V. (2006), Criterii de alegere a unei metode de sondaj în cercetarea statistică, *Analele Universității „Alexandru Ioan Cuza" din Iași*, Tomul LII/LIII, ISSN 0379-7864, pp. 373-379.
Jemna, D.V. (2005), Conceptul de reprezentativitate în cercetarea statistică, *Analele Universității „Alexandru Ioan Cuza" din Iași*, Tomul L/LI, 2004/2005, ISSN 124-516X, pp. 413-418.
Jemna, D.V. (2005), *Eficiența sondajului statistic*, Editura Sedcom Libris, Iași.
Jemna, D.V. (2015), *Sondajul statistic*, Editura Sedcom Libris, Iași.
Jemna, D.V. (2017), *Econometrie cu aplicații în R*, Editura Universității „Alexandru Ioan Cuza" din Iași.
Kalton, G. (1983), *Introduction to survey sampling*, Sage Publications, Inc, Newbury Park.
Kiaer, N.A. (1895), Observations et experiences concernant des denombrements representatifs, *Buletin de l'Institut International de Statistique*, 9(2), pp. 176-178.
Kish, L. (1965), *Survey sampling*, John Wiley & Sons, New York.

Krishnaiah, P.R., Rao, C.R. (eds.) (1988), *Handbook of statistics, Volume 6: Sampling*, Elsevier Science Publishers, North-Holland, Amsterdam.

Lanke, J. (1974), *Some contribution to the theory of survey sampling*, Unpublished Ph.D. Thesis, University of Lund, Sweden.

Laplace, P.S. (1951), *A philosophical essay on probabilities*, Dover, New York.

Laplace, P.S. (1814), *Essai philosophique sur les probabilités*, M.V. Courcier, Paris.

Lebart, L. (1992), *La qualité de l'information dans les enquêtes*, Éditions Dunod, Paris.

Levy, M. (1975), *L'information statistique*, Editions du Seuil, Paris.

Levy, P.S. (2008), *Sampling of populations: Methods and applications*, Fourth Edition, John Wiley & Sons, Inc., Hoboken, New Jersey.

Litwin, M.S. (1995), *How to mesure survey reliability and validity*, Sage Publications, London.

Mairesse, J. (Ed.) (1988), *Estimation et sondages*, Éditions Economica, Paris.

McEachern, W.A. (2000), *Economics*, South-Western College Publishing, Mason, Ohio.

Meynaud, H.Y., Duclos D. (2007), *Les sondages d'opinion*, Éditions La Découverte, Paris.

Mihoc, G., Urseanu, V. (1977), *Sondaje şi estimaţii statistice*, Editura Tehnică, Bucureşti.

Mihoc, G., Craiu, V. (1976), *Tratat de statistică matematică*, Volumul I, Editura Academiei R.S.R., Bucureşti.

Mill, J.S. (1896), *Système de logique déductive et inductive*, Éditions Felix Alcan, Paris.

Monfort, A. (1997), *Cours de statistique mathématique*, Éditions Economica, Paris.

Morgenstern, O. (1972), *L'illusion statistique. Précision et incertitude des données économiques*, Éditions Dunod, Paris.

Morin, H. (1993), *Théorie de l'échantillonnage*, Les Presses de l'Université Laval, Sainte-Foy.

Mouton, J., Marais, H.C. (1990), *Basic concepts. The methodology of social sciences*, HRSC Press, Pretoria, South Africa.

Mueller, J.H. (1970), *Statistical reasoning in sociology*, Houghton Mifflin Company, Boston.

Murthy, M.N. (1967), *Sampling theory and methods*, Statistical Publishing Society, Calcutta.

Neyman, J. (1934), On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society*, 97, pp. 558-606.

Neyman, J. (1938), Contribution to the theory of sampling human populations, *Journal of the American Statistical Association*, 33, pp. 101-116.

Nicod, J. (1924), *Le probleme logique de l'induction*, Éditions Felix Alcan, Paris.

Norden, R.A. (1972), A survey of maximum likelihood estimation, *International Statistical Review*, 40(3), pp. 329-354.

Onicescu, O., Ştefănescu, V. (1979), *Elemente de statistică informaţională cu aplicaţii*, Editura Tehnică, Bucureşti.

Onicescu, O. (1971), *Principes de logique et de philosophie mathématique*, Editura Academiei R.S.R., Bucureşti.

Pandurang, V.S. (1954), *Sampling theory of surveys with applications*, The Iowa State College Press, Ames.

Pasquier, A. (1969), Eléments *de calcul des probabilités et de théorie des sondages*, Éditions Dunod, Paris.

Pearson, K. (1912), *La grammaire de la science*, Éditions Alcan, Paris.

Plaisent, M. et al. (2009), *Introduction à l'analyse des données de sondage avec SPSS: guide d'auto-apprentissage*, Presses de l'Université du Québec.

Popper, K.R. (1991), *La connaissance objective*, Éditions Aubier, Paris.

Popper, K.R. (1981), *Logica cercetării*, Editura Didactică şi Pedagogică, Bucureşti.

Porojan, D. (1993), *Statistica şi teoria sondajului*, Casa de Editură şi Presă „Şansa" S.R.L., Bucureşti.

Quetelet, A. (1835), *Sur l'homme et le developpement de ses facultés, essai d'une physique sociale*, Éditions Bachelier, Paris.

Quivi, R., Van Campenhoudt, L. (1995), *Manuel de recherche en sciences sociales*, Éditions Dunod, Paris.

Rao, P.S.R.S. (2000), *Sampling methodologies with applications*, CRC Press, London.

Rao, J. N. K. (1969), Ratio and regression estimators. In N.L. Johnson, & H. Smith (Eds.), *New development in survey sampling*, pp. 213-234, Wiley, New York.

Sampath, S. (2001), *Sampling. theory and methods*, CRC Press, London.

Saporta, G. (1990), *Probabilités. Analyse des données et statistique*, Éditions Technip, Paris.

Sayer, A. (2000), *Method in social science*, Routledge, London.

Scheaffer, R.L., et al. (2012), *Elementary survey sampling*, Brooks-Cole Publishing, Boston.

Smith, J.G., Duncan A.J. (1945), *Sampling statistics and applications. Fundamentals of the theory of statistics*, McGraw-Hill, New York.

Stahl, H. (1974), *Teoria şi practica investigaţiilor sociale*, Editura Ştiinţifică, Bucureşti.

Stopher, P.R., Mayburg, A.N. (1979), *Survey sampling and multivariate analysis for social scientist and engeineers*, Lexington Books, Massachusetts.

Stuart, A. (1962), *Basic ideas of scientific sampling*, Charles Griffin, London.

Sudman, S. (1976), *Applied sampling*, Academic Press, New York.

Tabak, J. (2004), *Probability and Statistics: The science of uncertainty*, Facts on File, Inc., New York.

Tassi, P. (Ed.) (1987), *Les sondages*, Éditions Economica, Paris.

Tassi, P. (1989), *Méthodes statistiques*, Éditions Economica, Paris.

Tassi, P., Legait, S. (1990), *Theorie des probabilités en vue des applications statistiques*, Éditions Technip, Paris.

Thompson, S.K. (1992), *Sampling*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Tille, Y. (2001), *La théorie des sondages. Échantillonnage et estimation en populations finies*, Éditions Dunod, Paris.

Tiron, M. (1972), *Teoria erorilor de măsurare şi metoda celor mai mici pătrate*, Editura Tehnică, Bucureşti.

Wolter, K.M. (1984), An investigation of some estimators of variance for systematic sampling, *Journal of American Statistical Association*, 79(388), pp. 781-790.

Yates, F. (1953), *Sampling methods for census and surveys*, Charles Griffin, London.

Yule, U.G., Kendall, M.C. (1969), *Introducere în teoria statisticii*, Editura Ştiinţifică, Bucureşti.

Warwick, D., Lininger, C. (1975), *The sample survey: theory and practice*, McGraw-Hill, New York.

Zamfir, C., Vlăsceanu, L. (1998), *Dicţionar de sociologie*, Editura Babei, Bucureşti.

În aceeaşi colecţie au apărut:

Management. Fundamente, studii de caz şi metode econometrice,
Ioan Ciobanu

Performanţa şi eficienţa activităţii bancare,
Alin Marius Andrieş

Moneda, creditul bancar şi ciclurile economice,
Jesús Huerta de Soto

Dezvoltarea aplicaţiilor orientate obiect pe platforma Java,
Cătălin Strîmbei

Diversitate culturală în management. O abordare interregională,
Angelica-Nicoleta Onea

Competitivitate industrială,
Petrică Corăbieru

Uniunea Europeană şi competiţia globală,
Roxana Paraschiv

Influenţa timpului asupra evaluării afacerilor. Analiză. Diagnostic. Evaluare,
Ioan Dumitrean

Costurile de tranzacţie – o abordare instituţionalistă,
Andreea-Oana Iacobuţă

Activitatea bancară şi integrarea monetară europeană,
Vasile Cocriş, Elena Sireteanu, Alin Marius Andrieş

Performanţă şi risc în afaceri. Concepte, metode, aplicaţii,
Silvia Petrescu

Auditul financiar, de la normele naţionale la standardele internaţionale,
Ionela-Corina Chersan

Performanţa în serviciile de sănătate publică,
Maria Viorica Bedrule-Grigoruţă (editor)

Diagnostic intercultural. Competitivitate organizaţională prin mixare culturală şi despre creşterea performanţei manageriale prin sinergie interculturală,
Dumitru Zaiţ

Schimburile comerciale internaţionale între teorie şi realitate,
Liviu-George Maha

Contabilitatea activelor şi datoriilor financiare la societăţile comerciale,
Maria Carmen Huian

Echilibrul extern în economia deschisă,
Lăcrămioara Juverdeanu

Pieţe de capital,
Carmen Corduneanu, Laura Raisa Miloş, Claudiu Boţoc

Impactul politicii de finanţare asupra dezvoltării durabile a întreprinderii. O abordare financiar-contabilă,
Mihai Carp

Finanțarea firmelor în economiile emergente,
Silviu Ursu

Dinamica economică în viziunea instituționalistă,
Oana-Ramona Socoliuc

Dinamica piețelor de capital emergente,
Delia-Elena Diaconașu

Regimurile cursului de schimb în contemporaneitate. Perspective neoliberale,
Iulian Ihnatov

Monetary, Banking and Financial Issues in Central and Eastern EU Member Countries: How Can Central and Eastern EU Members Overcome the Current Economic Crisis? (vol. I), **Angela Roman, Sorin Gabriel Anton** (editors)

Monetary, Banking and Financial Issues in Central and Eastern EU Member Countries: How Can Central and Eastern EU Members Overcome the Current Economic Crisis? (Vol. II), **Angela Roman, Sorin Gabriel Anton** (editors)

Monetary, Banking and Financial Issues in Central and Eastern EU Member Countries: How Can Central and Eastern EU Members Overcome the Current Economic Crisis? (Vol. III), **Angela Roman, Sorin Gabriel Anton** (editors)

Diversitate culturală în management. O abordare interregională,
ed. a II-a, **Angelica-Nicoleta Onea**

Probleme actuale de economie publică,
Livia Baciu, Andreea Iacobuță, Alina Botezat, Mihaela Ifrim

Contabilitate în administrația publică,
Iuliana Georgescu, Leontina Păvăloaia

Gestiunea riscurilor financiare. Abordări teoretice și studii de caz,
ed. a II-a, **Sorin Gabriel Anton**

European Financial and Monetary Integration. Challenges of the Single Currency,
Angela Roman, Irina Bilan (editors)

Capitalul uman în noile democrații,
Cristian C. Popescu

Putere și globalizare,
Cristian C. Popescu
Lecții de economie financiară. Cum și de ce investim?,
Cristian C. Popescu

Auditul intern, de la teorie la practică,
Ionela-Corina Chersan, Cristina-Ionela Precob

Economiile de aglomerare și competitivitatea regională,
Raluca Irina Clipa

Demografia României,
Dănuț-Vasile Jemna

Econometrie. Cu aplicații în R,
Dănuț-Vasile Jemna

Investițiile străine, capitalul uman și creșterea economică,
Laura Diaconu Maxim

TIPARUL EXECUTAT LA
IMPRIMERIA EDITURII UNIVERSITĂȚII
„ALEXANDRU IOAN CUZA" DIN IAȘI
700109 Iași, Pinului 1A, tel./fax 0232 314947
Apărut: 2018
Comanda: 487

Informații și comenzi:
www.editura.uaic.ro
editura@uaic.ro